THE UNIVERSITY *of* EDINBURGH

Characterising and Measuring Human Episodic Memory

Iain Harlow

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2012

Abstract

Episodic memory, the ability to store and retrieve information from our past, is at the very heart of human experience, underpinning our identity and relationship with the world. Episodic memory is not a unitary phenomenon: in dual-process theory, researchers draw a distinction between familiarity, a rapid and automatic sense of oldness to a previously encountered stimulus ("I know that face"), and recollection, the reactivation of additional context from a particular episode ("We met at the York conference"). A fundamental objective in the study of human memory is to ground recollection and familiarity in neural terms. This requires accurately measuring the contribution of each from behavioural data, which in turn relies on an accurate characterisation of recollection. This thesis introduces a novel source retrieval task to demonstrate that recollection has two critical, and fiercely contested, properties: it is thresholded, i.e. it can fail completely, and successful recollection is graded, i.e. it varies in strength.

The consequences of this characterisation are explored. Firstly, familiarity and recollection are functionally separable retrieval mechanisms. Secondly, the models currently used to measure the contribution of each are generally flawed, and a corrected model is described which better fits, and explains, the extant data. Finally, the frequency of recollection is shown to be dissociable from its strength, a result which links behavioural data more strongly than before to a neurocomputational account of episodic memory, and which suggests a relationship between the representational overlap of memory traces and their retrieval.

This thesis necessitates a change in the way behavioural memory data is modelled, and consequently the interpretation of evidence underpinning neuroanatomical accounts of memory experience. Significantly, however, it also moves the field beyond a long-running debate and provides a deeper dual-process framework with which to address outstanding questions about the relationship between, and neural basis of, episodic memory processes.

Acknowledgements

Firstly, I would like to thank the many individuals who are involved in the Neuroinformatics DTC in Edinburgh and the Psychological Imaging Laboratory in Stirling, as well as the EPSRC and MRC for their financial support, and my family for their encouragement and everything else. Between them, they have all enabled me to produce this thesis, and have provided the ideal environment in which to acquire and pursue my interest in research.

My particular gratitude goes to Pat Ferguson and Catriona Bruce, who have taught me the priceless value of highly competent administrative and technical support.

I would like also to thank the members of the PIL, past and present, partly for the stimulating intellectual discussions we have had, but mainly for being a uniformly nice group of people who made it (nearly) always pleasurable to study for a PhD.

Thanks also go to Mark van Rossum for his help, advice and support, and to David Donaldson for his unwaveringly excellent supervision. His enthusiasm for both the science and the development of his students is evident in the success of the PIL, and I am profoundly grateful for the opportunity to learn from and work with him.

Finally, I would like to dedicate this thesis to Maria, who has made the past few years the happiest of my life - even the one spent thesis writing.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Iain Harlow*)

Contents

List of Figures

List of Tables

Life is all memory, except for the one present moment
that goes by you so quickly you hardly catch it going.
Tennessee Williams

Chapter 1

General Introduction

We are each a product of our memory. Our beliefs, skills, personalities, habits and sense of self are all formed and influenced, consciously or unconsciously, by past experience. Far from being only a remote link to the past, memory in its many forms is a crucial component of what it is to be a complex human being. Yet human memory can also seem mysterious, unpredictable and mendacious. Our experience of memory differs widely: why do we recall some moments, however trivial, with clarity and an avalanche of associated detail, and yet other times search in vain to place a familiar face? Understanding how human memory works is an important goal for research, especially in the context of ageing populations and an increasing burden of memory decline with age and disease.

1.1 Defining Memory

Memory, even restricted to the context of human psychology, is not a single, simple phenomenon. The term memory can be understood to refer to a wide variety of ways in which information from the past can be brought to bear on the present. Firstly, it is common to distinguish declarative from non-declarative, or procedural memory. Briefly, non-declarative memory comprises forms of memory which are not consciously available or retrievable, but which nevertheless influence our present actions: an example of this is the 'muscle memory' which allows us to navigate our environment, or to wield a tennis racquet. Non-declarative memory does not require explicit, conscious direction - indeed, it is often hindered by such

introspection. Declarative memory, in contrast, can be consciously retrieved, manipulated and acted upon. It is this form of memory which this thesis is concerned with.

Within declarative memory, many traditional models (e.g. Gobet, 1998, Roediger and Bergman, 1998, Tulving, 1972; though see also Squire and Knowlton, 1995) draw a distinction between semantic memory, which is memory for facts and knowledge about the world which is not tied to a specific episode ("European Elk are known as Moose in America") and episodic memory, the subject of this thesis, that is memory for individual moments, events or experiences ("I tasted Elk for the first time a few weeks ago, it was delicious"). It is immediately obvious that the distinction cannot be so cleanly made as a naïve interpretation of those models might suggest; semantic memories are presumably influenced by or based upon episodic memories, and it is difficult to determine a sharp boundary between the two types. Nonetheless, there is neural evidence for a dissociation between semantic and episodic memory; for example impairments in each system can arise independently of the other as a consequence of neurobiological damage (Mayes et al., 2001, Temple and Richardson, 2004, Vargha-Khadem et al., 1997).

The questions we pose in this thesis are related primarily to episodic memory, and in particular how these memories are retrieved. We note that recognition of individual episodes can be supported and influenced by forms of memory that are not episodic; semantic memory as we have already noted, but also priming (Olichney et al., 2000, Yovel and Paller, 2004). Non-mnemonic or metacognitive reasoning can also be used to judge whether a stimulus has been previously encountered, for example by calculating that "I would have remembered that clearly, therefore I did not encounter it". Primarily, however, episodic memory is described in terms of two main components: Familiarity and recollection.

1.2 Dual-process theory: Familiarity and recollection

We experience the retrieval of episodic memories in phenomenologically different ways. Often, we are able to retrieve information and details from an event, sometimes even to the extent that sensations and emotions we experienced at the

time are triggered and re-experienced in the present. Other times this detail and context seems absent, but we may still be able to (confidently) recognise objects as having been previously encountered. A familiar image might trigger a related event to be remembered, but even when it does not, we are often left with the certain knowledge that the image relates to a stored memory, albeit one which is temporarily inaccessible.

A single-process view argues that these experiences primarily reflect different strengths of a common memory retrieval process (Donaldson, 1996, Dunn, 2004). Stronger retrieval provides more and richer details, triggering a more vivid memory experience. An alternative dual-process view of memory holds instead that these reflect differential engagement of mnemonic processing (e.g. Joordens and Hockley, 2000, Onyper et al., 2010, Reder et al., 2000, Yonelinas, 1994). Most commonly, these two experiences are ascribed to recollection and familiarity (Yonelinas, 2002a). By this view, familiarity is typically described as a relatively rapid unidimensional signal that provides a quantitative measure of the likelihood that a stimulus has been previously encountered. New stimuli will feel, on average, less familiar than old stimuli (or indeed those which are sufficiently similar to old stimuli). Recollection, in contrast, is the qualitative retrieval of context, details, and other multidimensional information related to a particular event, and is often regarded as being slower and more effortful or consciously directed than familiarity. The definitions are at least partly derived from the separate subjective experiences of 'knowing' something has been encountered, and that of 'remembering' the context, a distinction which is classically illustrated by George Mandler's influential 'butcher-on-the-bus' phenomenon (p.253):

> Consider seeing a man on a bus whom you are sure that you have seen before; you "know" him in that sense. Such a recognition is usually followed by a search process asking, in effect, Where could I know him from? Who is he? The search process generates likely contexts (Do I know him from work; is he a movie star, a TV commentator, the milkman?). Eventually the search may end with the insight, That's the butcher from the supermarket!
>
> Mandler (1980)

Given their phenomenological basis, subjective introspection provides an intuitive understanding of the distinction between recollection and familiarity. Strong evidence for a dual-process view of memory, however, comes from elsewhere: Notably (but not exclusively) from amnesic patient studies, neuroimaging, and behavioural data. In these very different contexts, familiarity and recollection have been dissociated, supporting the dual-process view.

1.2.1 Behavioural differences between familiarity and recollection

Behavioural data have been used to dissociate recollection and familiarity along functional lines, by measuring the effects of different manipulations on recollection and familiarity. This approach has yielded a number of differential effects. Firstly, recollection appears to be more sensitive to conditions during encoding than familiarity is. Dividing attention at study generally has a greater negative impact on recollection than familiarity (Craik et al., 1996, Jacoby and Kelley, 1992, Yonelinas, 2001), as does the administration of benzodiazepines (Bishop and Curran, 1995, Curran et al., 1993, Hirshman et al., 2002). Later recognition is also known to be improved by deep (i.e. semantic) compared to shallow (i.e perceptual) processing of a word at encoding, and this 'Levels of Processing' manipulation has a greater effect on recollection than on familiarity (Gardiner et al., 1996, Khoe et al., 2000, Rajaram, 1993, Wagner et al., 1997). Finally, just as it does during study, dividing attention at test also disrupts recollection while leaving familiarity comparatively spared (Anderson et al., 1998, Dodson and Johnson, 1996, Troyer et al., 1999).

In contrast, other manipulations have the opposite effect, impacting familiarity more than they do recollection. For example, familiarity appears to be more sensitive to modality differences between study and test words (Gregg and Gardiner, 1994, Toth, 1996), though interestingly perceptual differences introduced between study and test for pictures lead to reductions in both recollection and familiarity (Yonelinas and Jacoby, 1995) or recollection alone (Rajaram, 1996). Familiarity is also affected by the retrieval bias adopted by participants during a recognition test. When participants adopt a more liberal response criterion (thereby accepting more items as old, see Section 2.2.3), the proportion of tri-

als which are accepted on the basis of familiarity increases, but the proportion of recollected trials remains constant (Gardiner and Gregg, 1997, Postma, 1999, Strack and Foerster, 1995, Yonelinas, 2001). This suggests that, for item recognition at least, the strength of recollection is relatively constant, and greater than that of familiarity. Increasing the delay between study and test (at least in the context of relatively short experimental delays, i.e. on the order of seconds and minutes) reduces familiarity more than recollection (Yonelinas, 2002a), though speeding response deadlines at test to below 1000ms impairs recollection but not familiarity (Benjamin and Craik, 2001, Toth, 1996, Yonelinas, 1994). Together these results imply that familiarity has different temporal characteristics to those of recollection: it is available faster, but is less long-lasting.

Briefly, it is worth noting that the relationship between familiarity and implicit memory remains a source of contention. While some argue that implicit memory, such as priming, can be separated from familiarity entirely (Tulving, 1985), others have suggested that familiarity may be supported by, based upon, or equivalent to, some forms of implicit memory (Mandler et al., 1986, Yovel and Paller, 2004). Given that priming participants with stimuli at test (for example, by flashing a word very briefly before presenting it for recognition, or displaying a sentence which leads participants to think of the test word) increases estimates of familiarity for both studied and unstudied words (Kinoshita, 1997, Lecompte, 1995, Rajaram, 1993), it seems likely that in some cases priming may contribute to familiarity, at least in so far as it is normally measured experimentally. This does not necessarily mean that implicit memory and familiarity should be equated. In other cases, some amnesic patients have shown impaired familiarity in the absence of measurable deficits to implicit memory (Moscovitch et al., 1993, Stark and Squire, 2000). It is possible that different types of signal underpin 'familiarity' over short intervals (where the signal appears to decay rapidly, as discussed above) and longer intervals (where a sense of familiarity may be elicited for a location or face encountered many months or years ago). We do not attempt to separate a familiarity signal from implicit memory using the behavioural data we report in Chapters 5–7, and indeed any estimates of familiarity there are as consistent with the influence of priming effects as they are with explicit memory (as is arguably also the case for the majority of behavioural studies in the wider literature). Thus, while we shall often refer to recollection and familiarity for

convenience, it should be understood that in these chapters the term familiarity could be used interchangeably with "familiarity, priming or any other evidence of prior occurrence not dependent on recollection". While this may appear to be a problem, it is one which is inherent to the field as a whole. An important future aim, beyond the scope of this thesis, is to determine whether different mechanisms or sources of evidence may lead to a feeling of familiarity, or whether familiarity can be distinguished as a separate phenomenon to priming, for example. Most crucially, as we shall demonstrate, recollection can be separately distinguished from either of these definitions of familiarity, supporting the dual-process view of episodic memory.

1.2.2 Neural differences between familiarity and recollection

Patient studies provide further evidence of a dissociation between familiarity and recollection, this time on the basis of different neural substrates. Damage to the hippocampus and surrounding temporal lobe has been repeatedly associated with severe deficits in recognition (Hamann and Squire, 1997, MacAndrew et al., 1994, Stark and Squire, 2000), suggesting that these structures are crucial to episodic memory. Notably, however, more focused damage to the hippocampus has been reported by many different groups of researchers as having a disproportionate effect on recollection, with familiarity being relatively spared (Aggleton et al., 2000; 2005, Baddeley et al., 2001, Bastin et al., 2004, Bowles et al., 2010, Holdstock et al., 2000, Jäger et al., 2009, Mayes et al., 2002; 2001, Peters et al., 2009, Quamme et al., 2004, Tsivilis et al., 2008, Vann et al., 2009, Vargha-Khadem et al., 1997, Yonelinas, 2002b, Zola-Morgan and Squire, 1986). In contrast, damage to perirhinal cortex has been associated with selective deficits in familiarity (Bowles et al., 2007), and patients with more widespread damage across both hippocampus and surrounding cortex show deficits in both familiarity and recollection (Knowlton and Squire, 1995, Schacter et al., 1996, Verfaellie and Treadwell, 1993, Yonelinas et al., 1998).

Lesion studies, uniquely, allow the function of a particular brain structure to be investigated by studying the effect of its removal. Thus, such data remain an irreplaceable source of evidence for a functional and structural dissociation between recollection and familiarity. Conclusions drawn from patient data are,

however, vulnerable to the charge that they may not generalise to the healthy brain. The brain is highly plastic, and long-term damage in particular may produce qualitatively different neural circuitry to compensate for loss of function. Furthermore, precise characterization of the extent of neural damage in each patient is very difficult, and is limited by the resolution and reliability of structural imaging techniques (Rempel-Clower et al., 1996). This means that a particular cognitive deficit might be associated with visible damage in a patient, or it might be caused instead by other damage which has simply not been identified. Patients are generally compared to healthy controls, whereas the more informative (but impossible) comparison would be pre- and post-insult comparisons within the patient population. Finally, the cohort being examined is often very small, and sometimes consists of a single participant, meaning that idiosyncratic strategies or qualitative individual differences would have a large effect on the results. With respect to this final point, studies involving larger numbers of patients, such as Vann et al. (2009) and Tsivilis et al. (2008), are particularly valuable.

Nonetheless, given the difficulties inherent to neuropsychological data, another important source of evidence for the dissociation of familiarity and recollection is functional neuroimaging performed on healthy participants. Two widely-used functional imaging techniques - electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) - have between them provided considerable evidence for a dissociation between recollection and familiarity. Event-Related Potentials (ERPs, formed from EEG) have comparatively low spatial resolution, and are therefore unable to provide compelling evidence for the distribution of neural substrates for familiarity and recollection. They do, however, allow the two types of retrieval to be qualitatively dissociated on the basis of temporal characteristics and scalp topography. We shall discuss in further detail some ERP effects which have been linked to familiarity and recollection in Chapters 8 and 9, but for now we note that several studies have found dissociations between the two (e.g. Curran, 2000, Curran and Cleary, 2003, Donaldson and Rugg, 1998, MacKenzie and Donaldson, 2007, Rugg et al., 1998; 2002), strongly supporting the view that familiarity and recollection are served by distinct neural substrates.

Evidence for which brain regions might comprise these substrates comes primarily from fMRI studies. To a large extent, these appear to broadly corroborate the picture gained from patient studies outlined above, supporting a link between

hippocampal activation and recollection (e.g. Davachi, 2006, Davachi et al., 2003, Montaldi et al., 2006, Ranganath et al., 2003; see Diana et al., 2007 for a review). It is important to note, however, that activation in a wide range of brain regions has been linked to episodic retrieval, not only medial temporal lobe cortex and hippocampus (Buckner and Wheeler, 2001, Henson, 2005).

1.2.3 Are familiarity and recollection really separable on the basis of neurobiology?

The true picture is of course likely to be more complex than a simple one-to-one mapping between memory experience and sharply defined brain regions. Nonetheless, as should be clear from the short review we have just undertaken, there is a large and relatively consistent body of evidence that familiarity and recollection - defined subjectively and in cognitive terms - do reflect genuine differences at both functional and neural levels. Just how clear-cut these differences are, and how different brain regions or networks support memory, are much more open and fiercely contested questions (see e.g. Cowell et al., 2010, Montaldi and Mayes, 2010, Ranganath, 2010, Shimamura, 2010, Wixted et al., 2010, Yonelinas et al., 2010; a selection of views on this topic which appear together, with others, in a recent issue of the journal *Hippocampus*).

Together with patient data (as well as lesion data and intracranial recordings from animals, and other evidence, see Ranganath, 2010, Rugg and Yonelinas, 2003, Yonelinas, 2002a for reviews) the results outlined here have led to the development of dual-process models which regard the hippocampus as crucial to recollection, but not to familiarity (Brown and Aggleton, 2001, Diana et al., 2007, Mayes et al., 2007, Norman and O'Reilly, 2003). This is by no means a universally accepted view, even within those who subscribe to dual-process theories, and others suggest in particular that the mapping of recollection onto the hippocampus is not so well supported by existing evidence as these models have argued (Manns et al., 2003, Squire et al., 2007, Wais et al., 2008, Wixted et al., 2010). These authors highlight evidence that familiarity may also be impaired following damage confined to the hippocampus (Cipolotti et al., 2001, Kirwan et al., 2010, Manns et al., 2003, Manns and Squire, 1999, Reed and Squire, 1997, Stark et al., 2002, Stark and Squire, 2003, Verfaellie et al., 2000, Wais et al., 2006), and that medial temporal

lobe cortex (particularly the perirhinal cortex) have been linked to the retrieval of associations (Eldridge et al., 2005, Jackson and Schacter, 2004, Kirwan and Stark, 2004). Some of these authors also propose that elevated hippocampal activity observed for 'recollection' conditions in fMRI studies may in fact reflect stronger, but not qualitatively different memory, from that which does not activate the hippocampus. In other words, the hippocampal/cortical separation which some models map on to familiarity and recollection may in fact map on to strong and weak memory, or a different mapping altogether (Wixted et al., 2010).

These broad positions perhaps represent the two main opposing views on the neurobiology of recognition memory, acknowledging that there are nuances on both sides as well as alternative ways of characterizing the topic (e.g. Henson and Gagnepain, 2010). In the remainder of this chapter and the next we highlight two key areas of debate, which we shall demonstrate are related, that may have profound effects on how strongly the extant data can be characterized as supporting the neurobiological separation of familiarity and recollection. Firstly: Under what circumstances, if at all, can familiarity support recognition of novel associations? Secondly: How valid are current assumptions used to estimate recollection and familiarity, and how significant an effect do these assumptions have on the conclusions ultimately drawn?

1.3 Associative Recognition Memory

So far we have described episodic memory purely in terms of encoding, storing and retrieving some piece of information from an experience. Inevitably, however, the nature of this information affects memory. In particular, a fundamental classification can be made between item memory and associative memory. While item memory allows the observer to recognise an individual stimulus and identify it as being previously encountered, it is associative memory which is used to place this item in context: to bring to mind all its spatial, temporal and abstract relationships with the rest of the world.

1.3.1 Why is associative recognition memory important?

As we shall soon see, accounting for associative recognition data is a major challenge for dual-process models of memory. Before we examine this practical relevance, however, it is also worth highlighting the importance of associative recognition in real-life. Memory for associations is a particularly crucial aspect of memory, which underpins our ability to construct and encode complex relationships from individual experiences. By extension, it is the retrieval of relationships which grounds our memories in the framework of 'episodes', and which ultimately permits interrogation of, and reasoning about, the world. Crucially, episodic associative memory allows relationships to be encoded, retrieved and used after a single encounter. This is of clear importance in everyday life, for example allowing us to recall a person's name and information about them, such as their occupation and where we know them from. Episodic associative memory also, however, plays a crucial role in the human 'experience'. A sense of self depends on the availability of episodes from our life, each defined by relationships between the components of those episodes and our role in them. Planning future actions, particularly in the abstract long-term, requires the mental manipulation of a model of the world - a model built up of the relationships between objects we have encountered.

This distinction between item and associative recognition can be drawn on philosophical grounds, but is it grounded on practical differences? The answer to this is undoubtedly yes. Firstly, the retrieval of relationships is functionally different from the retrieval of individual stimuli. While the latter can be supported by familiarity, and possibly also other forms of evidence such as priming, associative recognition is much more heavily reliant on the engagement of recollection (Donaldson and Rugg, 1998, Yonelinas, 1997) and does not appear to be as susceptible to priming effects (Cameron and Hockley, 2000, Westerman, 2001). Furthermore, the capacity of item memory is much greater than that for associations (Fagot and Cook, 2006, Standing, 1973, Voss et al., 2009), and may be spared even when associative recognition is damaged or lost (Holdstock et al., 2005, Westerberg et al., 2006).

A notable feature of recollection, and by extension memory for episodic associations, is its vulnerability to cognitive decline, both through normal aging or

diseases and disorders such as Alzheimer's (Healy et al., 2005, Howard et al., 2006, Jennings and Jacoby, 1997, Kopelman, 1989, Naveh-Benjamin, 2000). Such memory-related cognitive deficits can be debilitating, and their treatment and management will become an increasing burden in the future as life expectancies increase and populations age. Treatment can be improved with greater understanding of why recollection is damaged. Characterizing recollection in functional and cognitive terms, as well as the neural substrate which mediates it, is an important element in developing such an understanding. We address some fundamental characteristics of the recollection signal in Chapters 2 and 4.

1.3.2 Can familiarity support associative recognition?

As we have noted above, associative recognition has generally been considered to rely primarily on recollection (Hockley and Consoli, 1999, Yonelinas, 1997). Thus the finding that perirhinal cortex is implicated in the recognition of associations (Jackson and Schacter, 2004, Kirwan and Stark, 2004), is difficult to explain with an account describing familiarity as dependent on perirhinal cortex and recollection as reliant on hippocampus (Brown and Aggleton, 2001, Diana et al., 2007, Mayes et al., 2007, Norman and O'Reilly, 2003). One way in which the data can be reconciled with these theories is if in some cases associative recognition can be mediated by familiarity.

Mayes et al. (2007) and Diana et al. (2007) both adopt this argument, albeit in slightly different ways. In Diana et al. (2007), the authors appeal to the phenomenon of 'unitization', whereby associated components are perceived as a single item. The Binding of Item and Context (BIC) model outlined in this paper thus ascribes the function of the hippocampus specifically to storing and retrieving associations and relations between items (which are themselves encoded and retrieved primarily in medial temporal lobe cortex).

Critical to unitization is the generation of a novel concept or item which can be later recognised. For example, two words 'black' and 'mail' can together form the unitized compound word 'blackmail'. Note that the unitized representation does not in theory need to be related, semantically or otherwise, to either of its components; nor does it necessarily have to be a word. It is worth noting, however, that a majority of studies providing evidence of unitization have used

lexical stimuli (Bader et al., 2010, Diana et al., 2008, Ford et al., 2010, Giovanello et al., 2006, Graf and Schacter, 1989, Haskins et al., 2008, Opitz and Cornell, 2006, Quamme et al., 2007, Rhodes and Donaldson, 2007; 2008)[1]; a smaller number have found similar effects in faces (Jäger et al., 2006, Yonelinas et al., 1999). In contrast, a study using fractals determined that unitization was unlikely to have occurred to any great extent (Speer and Curran, 2007).

In sum, however, there would appear to be consistent evidence that under certain circumstances to-be-associated pairs might be unitized and then recognised on the basis of familiarity. This is potentially of great importance, given the vulnerability of recollection to failure, disease and cognitive decline, and it is worth investigating how the brain might be able to compensate for the loss of such a crucial function. Unitization also provides a way of reconciling extant data to dual-process models, such as BIC, which predict that perirhinal cortex activity should be linked to item rather than associative memory. The theory is disputed by some, however, who argue that while unitization may occur in specific circumstances it is unlikely to provide a general explanation for evidence of familiarity in associative recognition (Mickes et al., 2010).

An alternative view holds that while Mickes et al. are correct in their assertion that unitization does not generally occur, it is still possible for non-unitized pairs to elicit familiarity on a later recognition test. In Mayes et al. (2007) and Montaldi and Mayes (2010), pairs of items are hypothesised to be recognised on the basis of familiarity only when they share a stimulus class, or 'domain'. This domain dichotomy theory predicts that similar, closely related items will be represented by overlapping populations of neurons in pre-hippocampal cortex (so long as they are encoded simultaneously), and in particular in the perirhinal cortex. Here, cortical circuits are believed to form representations which can be later used to distinguish familiar from unfamiliar stimuli, and crucially these representations are thought to exhibit pattern-generalising properties (Norman and O'Reilly, 2003), causing increased familiarity signals for similar lures as well as previously studied targets (Hintzman et al., 1994). The similar (within-domain) pairs whose representations converge in perirhinal cortex can be distinguished from novel pairings on the basis of a familiarity signal arising from their combined representation in perirhinal cortex. Dissimilar (between-domain) pairs would instead not converge

[1]One of these, Diana et al., 2008, linked words with one of two colours in a source task.

until the hippocampus, where they will be bound by pattern-separating algorithms and require explicit recollection of their links to be retrieved as a pair. Data consistent with the theory have been found in a patient with relatively selective hippocampal lesions (Mayes et al., 2002) and hypoxic patients (Düzel et al., 2001, Vargha-Khadem et al., 1997); there is less clear evidence of domain dichotomy than unitization in healthy participants. The only study which, to our knowledge, explicitly makes this claim reports an associative recognition task in which a within-domain (face-face) condition showed apparently greater *reliance* on familiarity than a between-domain (face-name) condition (Bastin et al., 2010). While this result may be consistent with domain dichotomy, it is far from definitive evidence for two reasons. Firstly, the effect could have been caused by item type (introduction of lexical stimuli) rather than relationship type (domain) as claimed, since the components were not matched across conditions. Secondly, the effect observed (greater familiarity *relative* to overall performance) is consistent both with the claimed effect of increased familiarity and an alternative explanation of reduced recollection.

Despite different theories for how familiarity might contribute to associative recognition, evidence of its contribution has, as we have seen above, been reported by a number of different groups of researchers. It is possible, however, that much of this evidence may actually reflect recollection, and be misinterpreted as familiarity because of incorrect assumptions about how familiarity and recollection should be measured (Mickes et al., 2010, Wixted et al., 2010). In this thesis we shall investigate how associative recognition might be supported by familiarity and, in particular, whether the theories of unitization and domain dichotomy outlined above provide good explanations of why it may do so. Before we do so however, we shall examine more carefully the arguments over how to measure familiarity and recollection. As we shall see, the answer is not trivial and radically different conclusions may be drawn from the same data, depending on the approach taken.

1.4 Measuring recollection and familiarity

The functional and neural dissociations outlined earlier in the chapter, and the previous section's evidence of familiarity contributing to associative recognition,

rely on being able to accurately estimate the relative contributions of familiarity and recollection across conditions. One point of strength in this accumulated evidence is that a number of different methods are used to extract such estimates, ranging from participant introspection to calculation based on confidence ratings, and it can be argued that convergent evidence across different procedures provides a solid support for the dissociation in general (Yonelinas, 2002a, Yonelinas et al., 2010). Nonetheless, there is a strong argument that certain assumptions about how recollection is described in particular can have a significant effect on the interpretation of a dataset (Wixted, 2007a). This remains an important point of contention in the recognition memory literature, and has consequences in particular for the validity of the neurobiological separation of familiarity and recollection outlined above. The argument comes down primarily to whether or not recollection provides information of variable strength across trials, and whether it can fail completely or simply varies in strength from weak to strong. Given the importance of this point, we devote this section and the following chapter to the discussion of how best to model recollection and familiarity, and thereby estimate their contribution across different tasks.

One general approach has been to compare performance across different tasks, which presumably differ in the extent to which each relies upon recollection or familiarity, such as the process dissociation procedure (Jacoby, 1991), modified remember-know procedure (Mayes et al., 2007) or recall-recognition comparisons. One limitation of this approach is that precise estimates of recollection and familiarity are difficult to obtain, as the two tasks are unlikely to be process pure. Thus, it may not be possible to tell, given a difference across the two tasks, whether this difference can be isolated to an individual process or whether it simply reflects a quantitative difference (e.g. a greater effect on recollection, rather than a selective one) which does not necessarily imply a functional dissociation. Furthermore, having multiple tasks increases the risk that participants use differing strategies or criteria for each.

1.4.1 The Remember-Know procedure

A widely-used method for estimating recollection and familiarity within a single task is the Remember-Know (RK) procedure (Tulving, 1985). In this procedure,

participants supply an old or new response to a test item, but for old responses they additionally specify whether they 'Remember' the original presentation or simply 'Know' it[2]. Remember responses should in theory reflect greater recollection of the original episode, while Know responses should largely reflect familiarity in the absence of recollection (Mandler, 1980). The RK method has a crucial advantage over many other methods in that it allows individual trials to be allocated to a 'recollection' or 'familiarity' condition according to the response given. Comparisons between these two (selected) groups ought to have much greater statistical power than between two conditions which simply have (incidental) differing recollection and familiarity estimates. As a result, it is a widely used technique to separate trials for examination in neuroimaging experiments, which particularly benefit from increased effect sizes and statistical power. Another potential advantage if recollected trials are to be identified is that non-criterial recollection is detected and reported, though this might equally be seen as a disadvantage if the focus of the experiment is instead to track recollection of task-relevant information in particular.

The main disadvantage of RK judgments lies in their subjectivity. Remember-Know judgments are not process-pure (Wixted et al., 2010) and the relationship between remember/know and recollection/familiarity is not only unknown, but vulnerable to differences across both participants (e.g. due to individual interpretation) and tasks, researchers or labs (e.g. due to task instructions). Although there is evidence that in general, Remember and Know responses may differ qualitatively (e.g. Perfect et al., 1996, Wixted et al., 2010) it is possible that in some cases the two may simply reflect high and low confidence responses (Donaldson, 1996, Hirshman and Master, 1997). In fact, even when Remember and Know trials are matched in terms of confidence or accuracy (Montaldi et al., 2006, Wixted et al., 2010), if this occurs at ceiling levels it does not necessarily preclude differences in underlying memory strength. This is because trials assigned the maximum confidence rating may still differ widely in strength from each other, and very small error rates might either include comparatively significant proportions of motor error or unattended trials, or they might reflect weaker

[2]The procedure may vary slightly from this structure, for example a 'guess' option may also be included, or the old/new and remember/know decisions may be combined into a single remember/know/new decision. Additionally, there is some ambiguity about exactly how to interpret the results quantitatively, see for example (Mayes et al., 2007, Yonelinas and Jacoby, 1995).

memory. Finally, estimates drawn from remember/know judgments might be especially likely to misrepresent the contributions of familiarity and recollection if either is weak or infrequent - in such cases participants may feel obliged to provide a minimum number of responses in each category and consequently alter the mapping between experience and response accordingly. Chapter 5 in this thesis arguably provides just such an example, highlighting that RK is not well-suited to quantitative estimation of recollection and familiarity.

1.4.2 Sampling memory strength

Another limitation of each of the methods discussed above is that they provide relatively impoverished information about familiarity and recollection, producing at most an estimation of the proportion of hits supported by each. While this can sometimes provide sufficient data to separate two conditions, it is not necessarily a meaningful metric of each process, since it does not differentiate between the type of information retrieved, the frequency with which retrieval occurs or the strength of evidence it provides. The estimates derived can therefore be used, with the caveats outlined above, to investigate broad qualitative differences between familiarity and recollection (e.g. different neural correlates), but they cannot be easily used to distinguish between different descriptions of their functional characteristics (or isolate neural correlates beyond the familiarity/recollection distinction).

An alternative approach is to sample memory strength on each trial, for example by asking participants to rate their confidence, and fit these to a quantitative model. The data can be obtained from a single task, and does not rely on complicated instructions. Furthermore, the parameters drawn from such models have the potential to be quantitative and precise, and for a well-designed underlying model they should also be meaningful and accurate.

A major advantage of using confidence (or other ratings which reflect memory strength) is that the results may be informative about properties of the underlying signal. As we have noted, linking models across different levels - from neural circuits to observable behaviour for example - is a crucial aim in the effort to gain a more complete understanding of episodic memory. Quantitative models at different levels may be directly related to each other. Most relevantly in this

thesis is a possible link between patterns of memory strength derived from neural network models and those observed in behavioural studies (Elfman et al., 2008). Finally, such quantitative models allow more precise and detailed discussion and investigation of the processes they are assumed to reflect. As an example, we shall highlight a potentially important dissociation between the strength of recollection and its frequency, a distinction which is generally overlooked by the methods we describe above.

A further, more subtle, advantage of recording memory strength ratings is that the raw data can be reanalysed at a later date, and conclusions altered accordingly, as old models are superseded by more accurate ones. For example, data published by Yonelinas (1999) was reanalysed by other researchers (Slotnick and Dodson, 2005), who drew different interpretations as a result of the model used. Similarly, in Chapter 5 of this thesis, conclusions originally published using one model are reinterpreted using a more sophisticated version, on the basis of new evidence.

1.5 Key questions for this thesis

While a complete understanding of memory is an ambitious and long-term aim, in this thesis we shall address a number of key questions, the answers to which may illuminate some of its fundamental characteristics. Firstly, how should recollection be functionally characterised? There is fierce debate over whether recollection provides continuous strength, or is a probabilistic, state process. We directly address this question in Chapter 4, where we find evidence of a threshold: recollection is probabilistic. It can sometimes fail, and this (not weak memory) is what causes the experience of searching vainly for some detail of an event.

Secondly, how do familiarity and recollection support the crucial ability to remember associations and relationships? While recollection was previously believed to be crucial for associative memory, recently considerable attention has been focused on circumstances in which familiarity might also contribute to the recognition of previously-learned associations. In Chapters 5–7 we investigate this issue, finding that recollection may in fact be more crucial in most circumstances, and that the role of familiarity can be exaggerated by some current methods of

measuring memory properties.

Thirdly, therefore, how should recollection and familiarity be measured? The results from Chapter 4 suggest that some widely used models may be critically flawed in the assumptions they make about recollection, and we argue for an alternative approach which also allows the detection and investigation of other potentially significant patterns in the strength of memories. We use this model, combined with electrophysiological imaging, to examine how associative memory is supported by different memory processes (Chapters 5–9). We also address some ways in which memory is affected by properties of the stimuli being remembered, with corresponding implications for how stimuli are represented in the brain.

This final question - how should recollection and familiarity be measured? - is of particular practical importance in the field of episodic memory research. Many quantitative models have been suggested over the past fifty years to address this question, and debate as to which provide an accurate picture of recollection and familiarity - and therefore reliable and useful evidence in memory studies - is fierce and ongoing. In the next chapter we shall review this debate in some detail, and highlight the critical open questions which will motivate the first empirical chapter of this thesis.

Chapter 2

Quantitative models of episodic memory

One of the major aims of investigating memory, and in particular episodic recognition, is ultimately to explain the human experience of remembering - including in neural and neurobiological terms. Thus, it is of critical importance to relate phenomena at one level of investigation, such as human behaviour, to properties at a different level, such as the activity across a given network of neurons. We experience a range of memory phenomena, from the vivid recollection of an emotional experience to a fleeting sense of déja-vu for the present, and to begin to understand these experiences in terms of underlying neural systems we must define and quantify psychological concepts and behaviour.

2.1 Approaches to modelling memory data

Several different approaches and levels exist at which models can be made. Broadly speaking, some very large scale global models attempt to characterise an entire set of complex cognitive processes in a coherent, qualitative framework. Examples include Tulving's model of declarative memory (Tulving, 1985), Mandler's model of recognition (Mandler, 1979) as well as, more recently, neuroanatomically-constrained models of episodic memory such as the BIC model (Diana et al., 2007), CRAFT (Montaldi and Mayes, 2010) or the Representational-Hierarchical framework (Cowell et al., 2010). The aim of such models is to draw together

research from across the field, carried out by independent researchers, into a coherent framework. This framework can then be used to clarify the broad conclusions of an area of research, generate testable predictions (once it is sufficiently specified), adapt to new evidence, and be used to guide future research directions. Such models are usually broad enough to encompass data from many different modalities and research areas, such as behavioural data, imaging such as PET, fMRI and EEG, animal models and lesion studies.

2.1.1 Mechanistic models

A complementary approach is to describe particular types of data in quantitative, theoretically grounded terms. Roughly speaking, the quantitative models used in the study of episodic recognition can be separated into two types, namely mechanistic and measurement models. Mechanistic models begin with some generally agreed principles and try to establish what (otherwise non-obvious) predictions those assumptions might lead to. For example, neurocomputational models often begin by assuming a particular network structure or retrieval algorithm in order to produce predictions that can be tested or compared to empirical data (Hasselmo and Wyble, 1997, Norman and O'Reilly, 2003, Treves and Rolls, 1994). Such models allow a better understanding of the mechanics of a particular phenomenon, which can then be placed in context and related to other phenomena via the large scale models described above.

Mechanistic models are particularly well suited to the problem of linking empirical data across different levels of investigation. For example, neurocomputational models attempt to map the connectivity and architecture of a neural network onto the pattern of responses expected in a given behavioural memory task. These models are constrained in one direction by the biological reality of the brain: neurocomputational models are often (though not always, see for example Greve et al., 2010) built to mimic the connectivity within and between particular brain areas, such as the hippocampus or neocortex (Bogacz and Brown, 2003, Hasselmo and Wyble, 1997, Norman and O'Reilly, 2003, Treves and Rolls, 1994). Being quantitative, these mechanistic models can be relatively tightly constrained in the other direction by carefully defining the patterns of behavioural, phenomenological or imaging data they are required to explain. Such patterns

can be defined and quantified using measurement models.

2.1.2 Measurement models

Measurement models, in contrast to mechanistic models, begin with a set of empirical data and attempt to describe it in theoretically grounded terms. The approach is to extract, from complex empirical data, some manageable lower-dimensional description, which clarifies the important aspects of the data that are of interest. So long as the parameters of the model reflect some meaningful aspect of the data, such models can thus be used to understand and interpret an observed dataset better by describing it in terms of these parameters. For example, signal detection models of episodic memory (outlined in detail later in this chapter) describe the data in terms of changing contributions of memory processes, or properties of the information being retrieved. As a result these signal detection models are capable of making mathematical predictions about such contributions or properties for different task conditions, stimuli or memory deficits. In assessing measurement models, it is important to recognise that their predictions are not made in a vacuum; indeed, it is worth stressing that the parameters of these models must be clearly defined theoretically in order to be useful. The reverse is also true; defining theories in quantitative terms helps to inform and constrain them, communicate them precisely and to yield and test predictions in a rigorous way.

An important aspect of measurement models in particular is parsimony. We should immediately acknowledge that such models are a tool to quantify and test our explanations and not an explanation in themselves. Complex phenomena such as the encoding and retrieval of information in the brain can only ever be crudely approximated by measurement models derived from simple, noisy data such as memory task responses. Accordingly, the fit of a model can always be improved by increasing its flexibility, most commonly by adding parameters but also by making the existing parameters more flexible. Increasing flexibility allows the model to explain more data, but it comes with a cost. The accuracy of parameter estimates will be reduced since there is less information available per parameter to estimate its value. Ultimately, model selection is a trade-off between three main properties: 1) Goodness-of-fit, i.e. how much of the observed data

can be explained, 2) Parsimony, i.e. how efficient is the model and how reliable are its parameter estimates and 3) Testability, i.e. how well grounded are the parameters in theory and how well do its predictions account for other types of evidence.

In this chapter, we shall focus on a series of important quantitative measurement models, which attempt to describe how behavioural responses in human memory relate to the type and strength of stored information. To a certain extent, these models and the assumptions on which they rest constitute the rules relating behavioural memory data to the underlying processes and neurobiology; the different assumptions made by each model are often fiercely and incisively contested as a result. We shall outline the evolution of the current set of competing accounts, and highlight the more controversial and consequential points of disagreement across the models. Nonetheless, these models have been proposed and assessed over a period of decades in some cases, and as a result there is broad consensus over certain aspects of them. In particular, the field of signal detection theory has been very successfully applied to recognition memory, and we now describe the signal detection framework upon which these models are based.

2.2 Signal detection theory of memory

Signal detection theory (SDT) is a theoretical framework describing how decisions about the presence or absence of a target signal are arrived at, based on the level of some varying underlying property. Since its development during the Second World War to analyse radar signals (Marcum, 1947), it has been successfully applied to a wide variety of contexts, from radiographic imaging (e.g. Goodenough et al., 1972) to the perception of sensory stimuli (e.g. Creelman, 1965), from where we can trace the start of its influence on the field of episodic memory.

At its simplest, SDT describes the conversion of analogue information into discrete responses by the interaction of two features of detection: the distributions of values taken by different classes along some continuum, and one or more decision criteria placed on this continuum. Generally, the values along this continuum carry information relevant to the decision being made. In a memory experiment for example, where the task is to identify previously studied target stimuli (the

signal) and reject unstudied lures, these values are often assumed to reflect evidence of oldness - greater values of which (by convention) indicate a greater probability that the stimulus of interest has been previously studied. Decisions about each stimulus are made by comparing their value to a decision criterion, and classifying it according to whether it is higher or lower than this criterion value. In the old/new task for example, items with values greater than the old/new criterion will be accepted as 'old' and those with lower values rejected as 'new'.

In practice, the position of a criterion can be freely varied by the decision maker, and its position determines the bias of responses. Adopting a low criterion means that more stimuli will exceed it, and therefore be accepted. This is termed a liberal bias. Correspondingly, a very high threshold will be exceeded by few stimuli, and is termed a conservative bias. The most liberal or conservative criteria will respectively accept or reject all stimuli regardless of any information about their class - in both of these cases performance is, by definition, at chance. The 'best' decision criterion lies somewhere between the two, and its position is contingent on the underlying distributions of targets and lures, as well as the definition of 'best', which in turn depends on the consequences of the decision. For example, a criminal court requires a high burden of evidence for prosecution (i.e. a high, conservative criterion is placed on the evidence) since falsely convicting an innocent person of a crime is considered to be a more negative outcome than wrongly acquitting a guilty person.

2.2.1 Evidence distributions

In contrast to the criteria, the underlying distributions of values along an evidence axis are not normally freely variable by the decision maker. For example, memory strengths of stimuli in an old/new task are influenced by many factors, including encoding conditions and the cues available at retrieval, but at the point of making the decision the values can be considered to be a fixed property of all of these factors. When examining behaviour, such as a memory task response, it is important to distinguish between the act of making a decision and the evidence on which such a decision is based. Signal detection theory allows these two properties to be separately investigated, meaning for example that memory theories can be tested on measures which are independent of the decision process.

As we shall discuss in this chapter, signal detection theory provides an excellent description of how responses to a stimulus relate to a given memory strength distribution (i.e. how a decision is reached) and has been widely adopted by researchers in the memory literature. The signal detection framework, however, tells us nothing about the distributions themselves. A considerable effort in recent memory research has therefore been devoted to modelling these distributions, with the aim of explaining how those observed for different classes of stimuli arise. In this chapter we outline three types or groups of models and in particular their interpretations in terms of cognitive phenomena such as recollection and familiarity. To do so, however, we must first define precisely what these models describe: what exactly is the evidence in a memory task?

Primarily, memory researchers wish to describe the distribution of signals relating to memory. Decisions are unlikely to be entirely based on memory strength, however; any information which is diagnostic for the decision being made can be taken into account. For example, participants in a memory experiment with equal numbers of targets and lures might consider an item more likely to be a lure if they know that more of the preceding trials were targets. With this caveat, in a carefully designed memory task (i.e. one designed to maximise reliance on memory as opposed to non-mnemonic evidence), the bulk of evidence is generally interpreted as relating to some underlying memory signal, often termed 'memory strength' (Wickelgren and Norman, 1966).

To properly interpret signal detection theory in a neural or cognitive context, it is important to understand what this memory strength (especially for unstudied items) actually means. The most common way in which it is interpreted is that every possible stimulus elicits some signal in the brain which is related to prior experience. This relationship may be direct, with the signal strength increasing as a stimulus is repeated and decreasing as a result of interference and forgetting (Wickelgren and Norman, 1966), or it may reflect a complex computation of the relative likelihood that a given stimulus has been previously encountered (Bernbach, 1967). This particular signal is defined by its accessibility, i.e. behaviour depends directly upon it, but it is likely itself to be a spatial and temporal function of signals across the brain. There have been attempts to characterise these signals more specifically, for example in terms of their neurobiology (Brown and Aggleton, 2001) or neural computation (Bogacz and Brown, 2003, Hasselmo and

Wyble, 1997, Norman and O'Reilly, 2003). Nevertheless, as in other fields such as perception (Billock and Tsou, 2011) their properties are likely to be the subject of much future investigation, the results of which may be one important part of a framework linking neuroscience and cognitive science more closely. Finally, we stress again that formally, the evidence distribution is precisely that and no more: a measure of evidence for or against each possible decision (Pastore et al., 2003, Wickens, 2002). What this evidence actually reflects can only be inferred by assuming further knowledge about what each decision is based on, for example by carefully designing a task so that only certain properties of a stimulus are diagnostic.

While these signals are therefore not defined precisely at a neural level by investigators using SDT, they are certainly subject to some broad assumptions. Firstly, the signals are assumed to relate in some way to previous experience, either directly or by an overlap with previously encountered stimuli. Secondly, they are believed to carry noise; they do not only reflect previous experience diagnostic to the decision being made. Thirdly, they are generally modeled as being (at the level of behaviour) unidimensional. In other words, the finely-grained, multidimensional neural activity elicited in response to a stimulus can be reduced to a unidimensional 'evidence' axis[1] corresponding to the common human experience of the signal: the sense that something is 'old' without being able to further subdivide or explain the feeling. Notably, from a dual-process perspective this evidence is not a fixed property until retrieval, since at that point different processes (such as those underlying recollection or familiarity) can recover information about previous experience, and these processes may vary in diagnostic strength (see Greve et al., 2010 for an explicit instantiation of this). Finally, the signals are often (though not always, see for example Shimamura, 2010) considered to represent a relatively complex sum of underlying activity or evidence, resulting in an approximately normal distribution of values[2].

[1] Although the evidence scale is commonly described as some unidimensional continuum of values there is no reason at all to expect that memory strength underlying it should not be represented by a multidimensional space, and several researchers have made efforts to expand the continuum to, for example, two dimensions (Glanzer et al., 2004, Hilford et al., 2002, Macmillan and Creelman, 2005, Rotello et al., 2004). Ultimately, however, multidimensional evidence is assumed to be reducible to a unidimensional decision axis in this space (Swets et al., 1961).

[2] Most models assume roughly gaussian distributions of memory strength as a default, even when explicitly acknowledged that this is largely a convenient approximation (Wixted, 2007a, Yonelinas et al., 2010). Others, however, have argued that memory strength should be modelled

The simplest manifestation of signal detection theory arising from these assumptions is the equal variance signal detection (EVSD) model (Parks, 1966, Wickelgren and Norman, 1966), which we outline below. Although this model is not now often used to describe memory strength distributions itself, the models we outline later in the chapter can all in some way be considered to be an extension or modification of this basic EVSD framework. It is therefore worthwhile introducing the EVSD model at this stage, and using it to illustrate the relationship between memory strength and behaviour, before introducing the more complex models which are the subject of the remainder of this chapter.

2.2.2 The Equal Variance Signal Detection model

The Equal Variance Signal Detection model is illustrated in Figure 2.1. In the context of an old/new decision, each stimulus has a single value on a unidimensional scale, and it is this value which is accessible to the decision maker. The population of unstudied (lure) values is normally distributed, with a mean value of zero and a standard deviation of 1 by convention. The population of studied (target) items is also normally distributed, with the same variance, but a mean > 0. In other words, the two populations are equivalent except that studied items have, on average, greater memory strength than unstudied items. To convert these values into, for example, an old or new decision, a criterion is applied. Targets which exceed the criterion are correctly accepted as old (hits) while those falling below it are misidentified as new (misses). Lures which exceed the criterion are misidentified as old (false alarms) and those below it are accurately characterised as new (correct rejections).

Within the EVSD model, accuracy on a task is given by the proportion of trials which are either hits or correct rejections, and is therefore determined not only by the distribution of targets and lures but also by the position of the decision criterion. Thus accuracy reflects a combination of both the information about the past that is available, and the way in which this information is used to produce a decision. The same underlying memory strength distributions could give rise to

quite differently, as a skewed distribution (Shimamura, 2010). The difference may be significant, since a normal distribution implies a large sum of information or activity, while a nonlinear or skewed distribution may imply that some sources of evidence, be they neurons, processes or brain regions, are particularly diagnostic and contribute more strongly to the decision.

a range of accuracies, simply as a function of the bias adopted at retrieval.

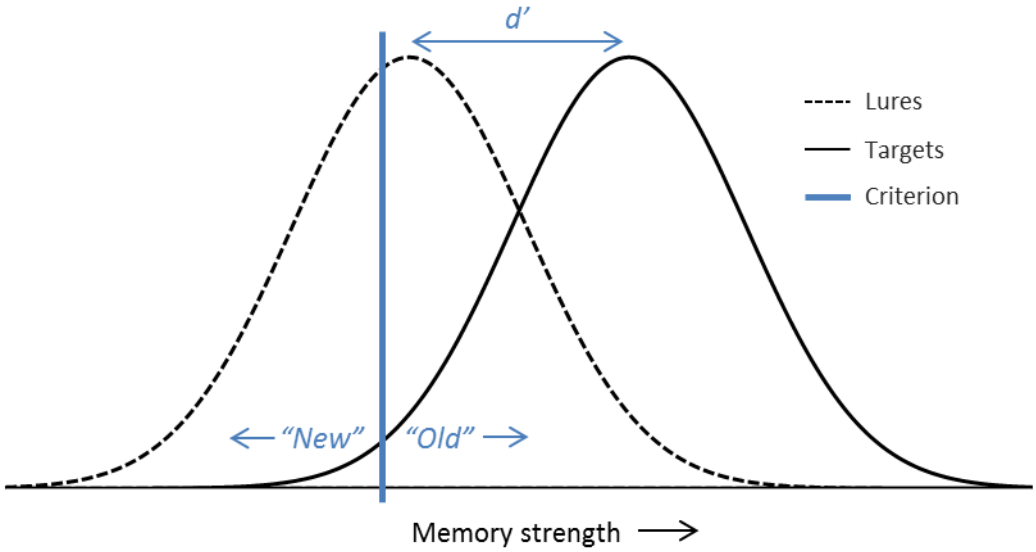


Figure 2.1: The equal variance signal detection model of episodic memory. The memory strengths of unstudied lures and studied targets each follow normal distributions, with matched variance, but targets have a higher average memory strength than lures. This is quantified by the separation of the two distributions in standard deviation units, denoted d'. Each memory decision is made by comparing the value of a given test trial to a criterion on the strength axis: trials with greater memory strength than the criterion are reported as 'old' and those with lower strength are reported as 'new'.

As a result, in signal detection theory performance is often characterised instead by the separation of the target and lure distribution means in standard deviation units. This value is known as discrimination and is denoted d' (Figure 2.1). Unlike accuracy, discrimination is theoretically independent of the position of the decision criterion, making it a more appropriate statistic to use when the subject of investigation is memory strength itself and not the way in which it is interrogated. For this reason, throughout this thesis overall performance on memory tasks will normally be summarised in terms of discrimination (details of how we calculate discrimination are provided in Section 3.1.3).

2.2.3 Examining multiple criteria

To accurately calculate discrimination, or indeed to learn anything about the distributions of evidence supporting a decision, these distributions must be sampled at different points. One way to do this is to manipulate the bias used by the participant, for example by changing the rewards and penalties for accepting or rejecting each trial. When the reward for correctly accepting a trial is much higher than the penalty for falsely accepting a lure, participants should adopt a more liberal bias (maximising their reward). If on the other hand the penalties of falsely accepting a lure are much higher, participants should adopt a correspondingly conservative bias (minimising their penalty).

A major advantage of sampling by this method is that it can be applied to non-human subjects (Fortin et al., 2004, Sauvage et al., 2010; 2008). For example, Sauvage and colleagues (Sauvage et al., 2008) tested the memory of rats by requiring them to judge whether a combination of odour and physical material (e.g. sawdust) had been previously encountered (intact) or was rearranged. By varying the effort required to make a rearranged judgment together with the food payoff for each type of correct response they were able to sample points along each rat's memory strength distribution and draw conclusions about their memory performance in the same way as for human participants. This method is not without controversy, however; see (Eichenbaum et al., 2008) and (Wixted and Squire, 2008) respectively for arguments for and against this approach.

More commonly in humans, the distributions are sampled by asking participants to rate their confidence from 1 to N on each trial. To the extent that this rating reflects genuine memory strength, $N-1$ implicit decision criteria can be reconstructed after the experiment by locating them between each consecutive pair of unique confidence ratings, as illustrated in Figure 2.2. A primary advantage of using confidence ratings to sample memory strength distributions is its experimental simplicity. Confidence is an intuitive concept to most people, which means the same measure can be collected from a wide range of participants with varying deficits in memory. This simplicity also makes it an easy statistic to gather in psychology experiments, since complex instructions are unnecessary and judging confidence typically takes a single short response per trial. A second, more subtle advantage, is that participants effectively divide up the decision space

themselves, which combined with careful task instructions can result in a more even and therefore informative sampling of the distributions. Of course there are also disadvantages to using confidence ratings, chiefly related to the fact that they represent an indirect, metacognitive assessment of memory strength. These disadvantages are addressed in more detail in Section 2.3.5 and Chapter 4.

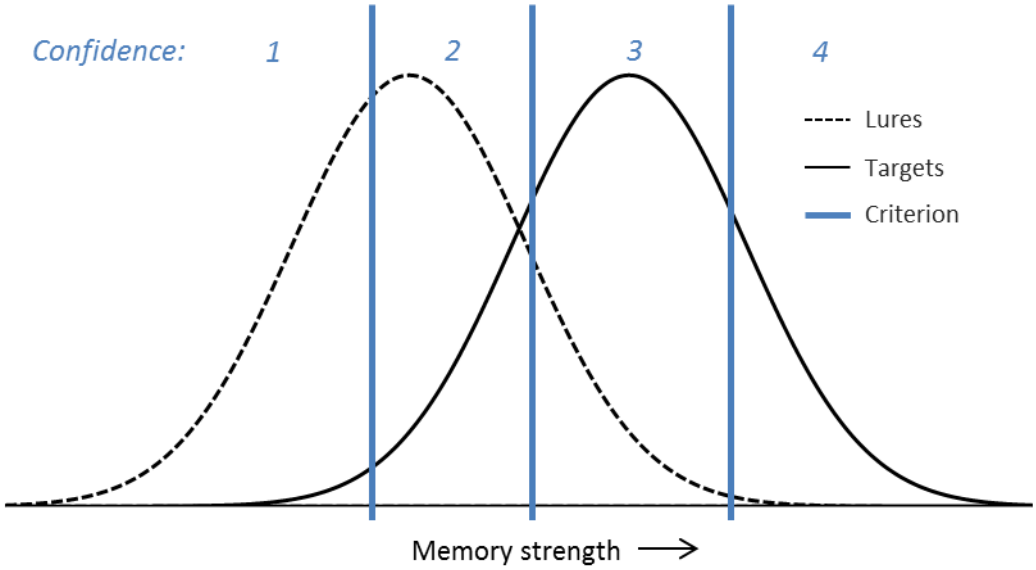


Figure 2.2: Reconstructing multiple decision criteria using confidence ratings. Each test item is rated according to how likely the participant considers it to be old, in this case on a scale from 1-4. Three decision criteria can be reconstructed by setting different minimum confidence ratings for an old judgment: 2 (most liberal), 3 or 4 (most conservative).

2.2.4 The receiver operating characteristic curve

Memory responses classified by multiple decision criteria can be used to calculate performance statistics, such as d', which are bias-independent. Beyond this, however, they can also be used to investigate the nature of the underlying memory strength distributions. One way to approach this is to examine the frequencies of hits and false alarms for decision criteria placed at different points along the distributions (Figure 2.3(a)). A Receiver Operating Characteristic (ROC) curve (see for example Fawcett, 2006) can then be formed by plotting each point in a two dimensional space, with the false alarm and hit rates on the x and y-axes

respectively (Figure 2.3(b)). For an individual participant and task, these points will lie upon a single line, the shape of which is informative since it reflects the distributions of lure and target memory strengths. Different models of these distributions therefore predict different ROC curve shapes, and so examining the shape of the ROC relative to these predictions has become a popular source of evidence for or against different episodic memory models (see Yonelinas and Parks, 2007 and Wixted, 2007a for reviews of this approach).

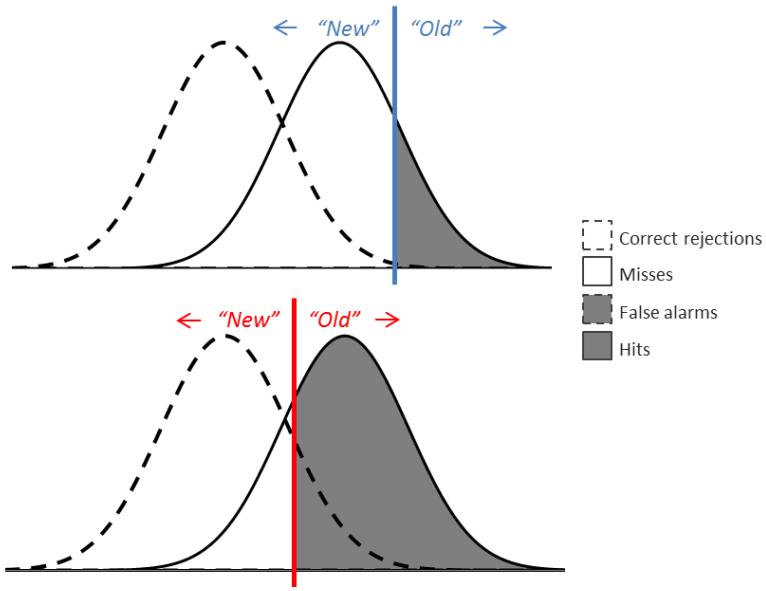
One reason why ROC curves in particular have become widely used is that they provide a concise visual representation of performance in a memory task across all possible levels of bias. As is clear from Figure 2.3, the patterns of underlying memory strengths for both targets and lures are transformed into a single curve in ROC space or z-ROC space. A researcher familiar with ROC curves can therefore use them to quickly and easily obtain a relatively complex summary of memory performance for a given task.

Two particular aspects of the ROC curve shape have been considered to be especially informative. Firstly, if the target and lure distributions are both Gaussian, the curve will form an inverse u-shape, i.e. it will be curvilinear. This property can be checked more clearly by plotting the data in z-space, as shown in Figure 2.4. If both target and lure distributions are normally distributed, the points will lie on a straight line; the method is similar in logic to using a QQ-plot to test for normality. In this manner the linearity of the z-ROC can be tested; any nonlinearities may constitute evidence against normally distributed targets or lures.

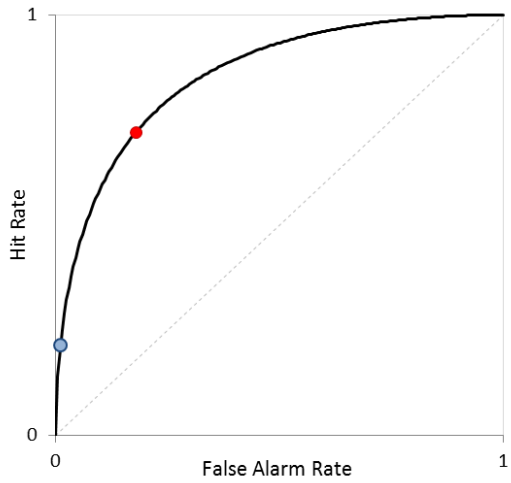
Furthermore, if targets and lures are matched in variance, as described by the EVSD model, the ROC curve will be symmetrical about a diagonal between $(0, 1)$ and $(1, 0)$. In this case, the z-transformed data will form a line with a gradient of 1. If, instead, the target variance is greater than the lure variance, the gradient of the z-ROC will be less than 1, and vice-versa. In sum, the z-ROC is a potentially powerful way of testing the assumptions of normality and equal variance.

2.2.5 Disadvantages of ROC and z-ROC curves

It should be noted, however, that there are a number of important disadvantages to using ROCs or z-ROCs alone to examine memory. In particular, since memory



(a) Conservative (upper; blue) and more liberal (lower; red) response criteria on the same memory strength distributions.



(b) The ROC curve resulting from the distributions sampled in (a).

Figure 2.3: The relationship between memory distributions and the ROC curve. Each possible response criterion (a) results in a particular false alarm and hit rate, corresponding to a single point in ROC space (b). Two such criteria and resulting ROC points are illustrated in blue and red. These points are constrained to lie upon a single 'ROC curve' (solid line in (b)), the shape of which is determined by the underlying memory strength distributions.

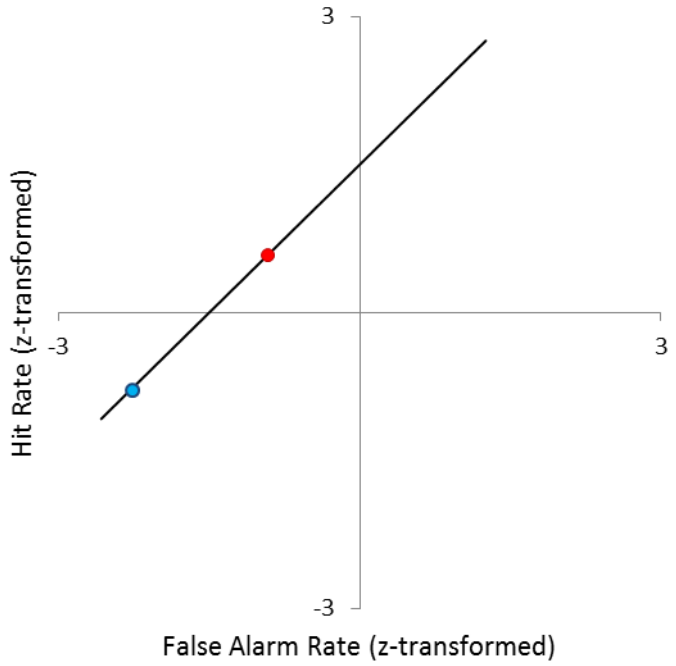


Figure 2.4: The z-ROC. The data from Figure 2.3 plotted in z-space. Since the underlying memory strength distributions are normal and of equal variance, the resulting z-ROC is a straight line with gradient 1, and has a y-intercept given by d'.

models tend to make predictions about ROC curves indirectly, via predictions about the memory strength distributions which give rise to them, fitting data to the ROC curve provides no clear advantage over fitting it to the underlying distributions directly. In fact, fitting data to the ROC curve can be problematic because when confidence is used to form ROCs, the points produced are not independent of each other. Since the data is collated at each point this also has the effect of exaggerating its signal-to-noise ratio; even noisy underlying data can produce a ROC which is apparently well interpolated by a smooth curve.

The fact that the data are not independent at each point also poses a problem for the normal approach used to examine the z-ROC. Hypotheses about the linearity of this curve are usually tested by regressing the set of z-transformed points against both linear and quadratic factors (e.g. Yonelinas, 1999), and the significance of the quadratic factor is assumed to reflect the significance of non-linearities in the z-ROC. The linear regression analysis assumes independence between points, however, an assumption which is not met by confidence-derived ROCs. There are a number of other reasons to be cautious about using this ap-

proach. Since the data can vary in both x and y directions, the analysis should be performed separately in each direction. More problematically, the quadratic model may not accurately describe the departures from nonlinearity present in the z-ROC; for example, an s-shaped curve might be captured by a cubic but not quadratic function, resulting in incorrect acceptance of the null hypothesis. Setting this problem aside, deviations from linearity may be hard to detect even when the underlying models are clearly non-Gaussian (Lockhart and Murdock, 1970), a problem which is compounded by the relatively low number of data points normally used to form ROC curves, and therefore their statistical power. Finally, some researchers have questioned the validity of confidence-derived ROCs altogether as a means of differentiating between different types of models (Bröder and Schütz, 2009, Malmberg, 2002). We shall return to this point in particular in more detail in Section 2.3.5.

2.3 Signal detection models

Despite the misgivings about interpreting ROC and z-ROC data mentioned above, signal-detection theory has nonetheless gained wide acceptance in the literature, in large part as a result of evidence from ROC analysis (for two excellent, competing, reviews see Yonelinas and Parks, 2007 and Wixted, 2007a). In fact, one final problem with using ROC curves to differentiate between models may, paradoxically, be a consequence of their widespread use: Although the models predict different shapes of ROC curve, such differences are not necessarily large in ROC or z-ROC space. The majority of models with broad support in the current literature fit most ROC data well and are often difficult to clearly separate using this method, leading to frequent disagreement over how to interpret the resulting data (for a notable example see Parks and Yonelinas, 2007 vs. Wixted, 2007b). Nonetheless, ROC analysis has been a useful tool to narrow down the field of plausible models, to the extent that, as we shall soon discuss, some consensus has arisen over certain aspects of them. In this section we examine how the EVSD account has evolved, leading to the current set of models.

2.3.1 The unequal variance signal detection model

The simple EVSD model (see Section 2.2.2) describes memory strength in terms of normal distributions, such that targets and lures have equal variance, but targets have greater average memory strength. Green and Swets (1966) proposed an update to the existing equal variance signal detection model, to account for the finding that targets could sometimes be associated with more variable memory strength than lures (Egan, 1958). In this unequal variance signal detection (UVSD) model, the distribution of strength values for targets is allowed to differ from that of lures, not only in terms of their means but also their variances (Figure 2.5).

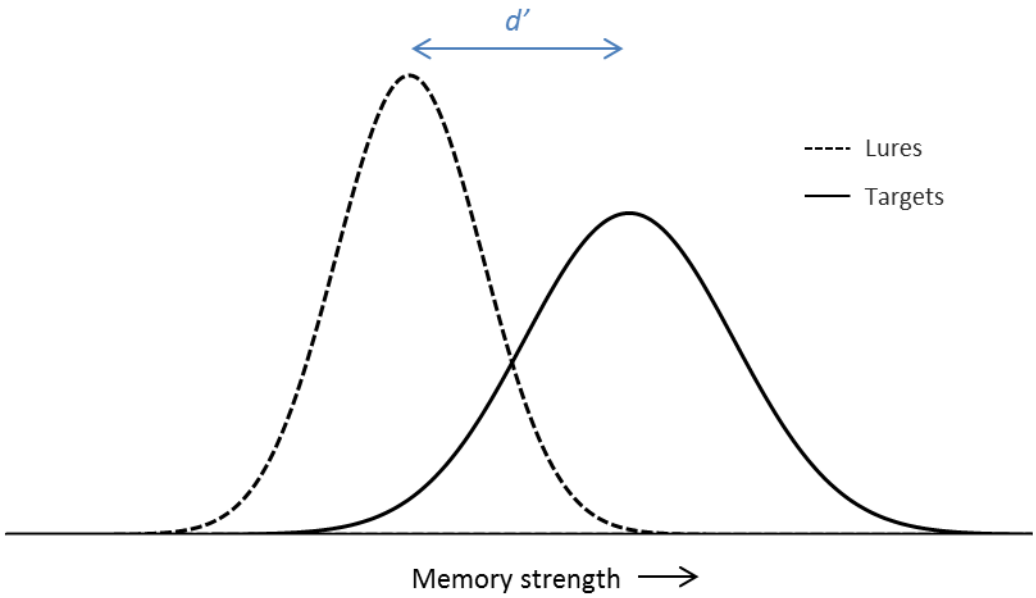


Figure 2.5: The unequal variance signal detection (UVSD) model. The model is identical to the EVSD model in Figure 2.1, except that the variance of the target distribution, $v(old)$, may differ from that of the lures (set to 1 by convention).

In the decades following the publication of (Green and Swets, 1966), many studies have reported evidence of greater target than lure variance in memory tasks, for example (Glanzer and Adams, 1990, Glanzer et al., 1999, Gronlund and Elam, 1994, Hirshman and Master, 1997, Ratcliff et al., 1994; 1992, Yonelinas, 1994), leading to a general rejection of the EVSD model in favour of the more flexible UVSD account. In fact, the UVSD model of recognition memory has had notable success in fitting empirical ROC data for this reason (Wixted, 2007a). However,

while the UVSD model can describe the differences in variance well, it does not automatically explain them. Indeed, Green & Swets (1966) themselves noted that (p.79):

> The justification for the Gaussian model with unequal variance is, we believe, not to be made on theoretical but rather on practical grounds. It is a simple, convenient way to summarize the empirical data with the addition of a single parameter.
>
> Green and Swets (1966)

Subsequently, others have sought to explain why greater target variance might arise theoretically, given that it has been so successful in describing the data from memory studies. It has been proposed that the variance arises as a result of adding different amounts of memory strength to targets (Hilford et al., 2002, Wixted, 2007a). By this 'variable encoding' formulation, studying a stimulus increases its memory strength, but critically, by some non-constant value. Assuming the strength added is normally distributed across studied items, the resulting distribution of target memory strength is a normal distribution with greater mean and greater variance than the unstudied lure distribution - i.e. the UVSD model. In this case, therefore, the UVSD model can be interpreted as being a single-process account, since the explanation of memory strength distributions does not require the invocation of more than one signal or retrieval process.

The UVSD model predicted by encoding variability has had some success in accounting for empirical data (Wixted, 2007a). Importantly, however, while encoding variability is compatible with a single-process view of recognition, the UVSD model itself is agnostic as to whether one or more processes are required for recognition, and therefore it is important to test the viability of encoding variability before declaring the UVSD model to be a single-process account. If accurate, encoding variability predicts that a set of stimuli studied for the same length of time each should vary less in terms of memory strength by retrieval than if they were encoded for different lengths of time, a prediction which was recently tested empirically (Koen and Yonelinas, 2010). The authors of this study, however, found that items studied for either 1s or 4s were no more widely distributed in memory strength than when each was studied for 2s.

An alternative or complementary account to encoding variability is provided by allowing the additional variance to arise because of a greater number of evidence sources for targets than lures. This sources-of-evidence view is more consistent with dual-process theory, a point stressed by Wixted and colleagues (Mickes et al., 2009, Wixted, 2007a, Wixted et al., 2010). From this perspective, recollection provides diagnostic information for targets, but not for lures, while a familiarity signal provides evidence for each type of stimulus. If all the available sources of evidence are combined to produce a single value for each stimulus, those with more sources of evidence will have greater variance in this value, for similar reasons to the variable encoding explanation above. Consistent with this explanation, amnesic patients with severe deficits in recollection, and who therefore should be reliant upon a familiarity signal which operates for both targets and lures, have been shown to produce relatively curvilinear and symmetric ROCs, with matched target and lure variance (Yonelinas et al., 1998).

In both cases, the explanation yields an expectation of the pattern of results across different tasks. In an old/new recognition study, targets should have a greater variance either because they are encoded with different strengths or because recollection of the original study episode provides an additional source of information. In a standard source retrieval task, however, where participants are required to recall some binary property of an original study episode (e.g. was a stimulus presented to the left or right of fixation), the UVSD model should collapse to the standard EVSD model. This is because both targets and lures will each be associated with a single study episode, so both encoding variance and sources of evidence are theoretically matched across the two.

Associative recognition is more complex, but targets (intact pairs) are associated with a single study episode, whereas lures (rearranged pairs) are associated with two episodes, either or both of which could be recollected. Thus, by either the variable encoding or the sources-of-evidence accounts, the total variance of memory strength for lures should exceed that for targets, which is a reversal of the pattern normally found for item recognition. Empirically, however, intact pairs tend to have a variance which is as great or greater than rearranged pairs (Healy et al., 2005, Mickes et al., 2010), making it somewhat difficult to interpret the functional meaning of the variance parameter in the UVSD model.

While item recognition tasks have often yielded quite linear empirical z-ROCs,

nonlinear z-ROCs have been found for associative recognition, source retrieval and some item recognition tasks (Rotello and Heit, 2000, Wixted et al., 2010, Yonelinas, 1997; 1999). These typically form a slight inverse u-shape, measured by a significant positive quadratic component, which suggests that either the lure or target distribution deviates from normality to some extent. The remaining models discussed in this section each posit a particular mathematical description of this deviation, as well as a functional interpretation of it.

2.3.2 The dual-process signal detection model

One of the earliest (and subsequently widely adopted) models was the dual-process signal detection or DPSD model (Yonelinas, 1994). In this model, recollection and familiarity are explicitly separated, with familiarity described by the existing equal variance signal detection model. Much like the sources-of-evidence interpretation of the UVSD model discussed above, increased target variance in item recognition studies is accounted for by the operation of recollection for previously-studied items only. Unlike UVSD, however, recollection is modelled as a thresholded, probabilistic process which can either succeed or fail. The simplest version of the DPSD model thus has 2 parameters: d', which describes the distance between the familiar and lure memory strength distributions in standard deviation units (identical to d' in the EVSD model) and $p(R)$, the probability of recollecting a target. Note here that $p(R)$ is a single parameter, not a function - it refers to the probability (between 0 and 1) of recollection, denoted R, occurring on each trial. The DPSD model is illustrated in Figure 2.6.

One advantage of the DPSD model is that it can be straightforwardly extended to associative recognition tasks. In an associative task, targets are intact pairs. The original study presentation for intact pairs can be recalled with probability $p(R)$, relabelled $p(R_a)$ here to clarify that it refers to the 'recall-to-accept' rate, i.e. recollection allows correct identification of intact pairs. Lures are rearranged pairs of previously-encountered items, and aspects of the original episodes in which either item was studied can also be recollected. The DPSD model therefore introduces a new parameter, $p(R_r)$, which is the probability that such an episode is recalled ('recall-to-reject'), thereby identifying the test pair as rearranged. Familiarity would operate in the same way as in item recognition, i.e. intact and

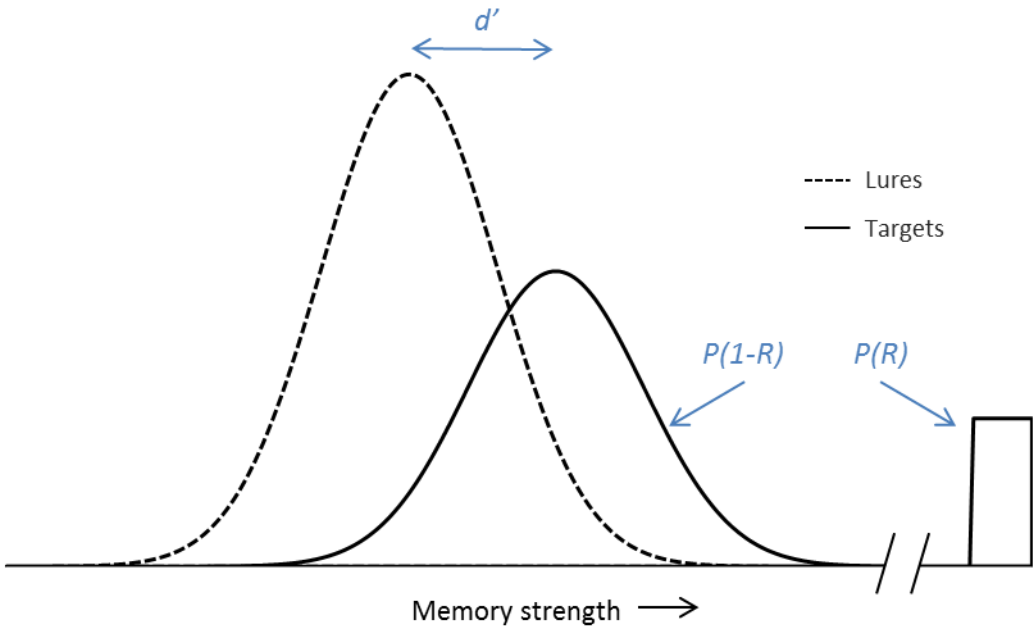
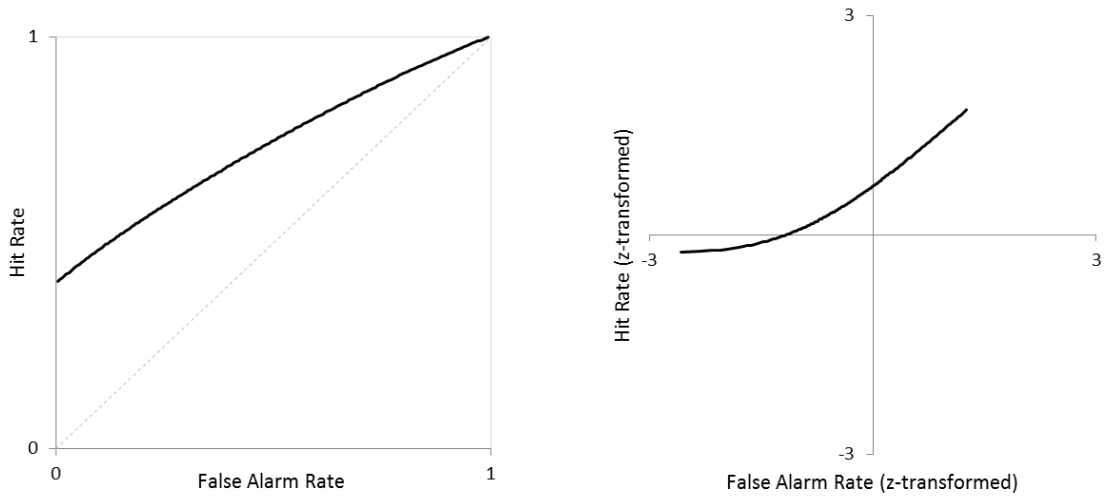


Figure 2.6: The dual-process signal detection (DPSD) model. The model assumes that some proportion $p(R)$ of target trials are recollected, and that these trials are always recognised as old with high confidence (illustrated by an arbitrarily high memory strength). The remaining $p(1-R)$ non-recollected targets are distinguished from lures using familiarity, which is described by an EVSD model.

rearranged pairs would have overlapping, equal variance normal distributions of memory strength, separated by d' standard deviations.

Normally, associative recognition would be expected to rely more heavily upon successful recollection than item recognition would, because the components of both intact and rearranged pairs have both been encountered and are, therefore, theoretically matched in terms of their familiarity. The familiarity of a pair of items would thus, on average, provide less diagnostic information for an intact/rearranged judgment than the familiarity of a single item would towards an old/new judgment. The DPSD model normally predicts ROC curves which are relatively curvilinear, but which intercept the y-axis at $(0,p(R))$ rather than $(0,0)$. When familiarity is less diagnostic, however, the model predicts a more linear ROC and correspondingly nonlinear z-ROC, Figures 2.7(a) and 2.7(b) respectively. Crucially, both associative and source recognition data fitting this prediction were published and replicated shortly after the model was proposed (Rotello and Heit, 2000, Yonelinas, 1997; 1999, Yonelinas et al., 1999).



(a) A relatively linear ROC predicted by the DPSD model with parameters $p(R) = 0.4$ and $d' = 0.2$.

(b) The corresponding z-ROC is u-shaped, not linear.

Figure 2.7: Linear ROC and nonlinear z-ROC predicted by the DPSD model. As d' approaches 0, i.e. as familiarity decreases, the DPSD model becomes more thresholded and predicts more linear ROCs, and nonlinear z-ROCs, than pure signal detection models (e.g. see Figure 2.3). Conversely, since the DPSD assumes that familiarity provides a Gaussian distribution of strength, when $p(R)$ approaches 0 the ROC becomes curved and the z-ROC linear, as for a pure signal detection model.

The DPSD model can be similarly extended to describe performance on a source task, where participants are required to remember some specific context of an episode, for example whether a previously studied item was spoken in a male or female voice, or in red or blue text. As in associative recognition, two separate recollection parameters are included, in this case to define the probability of recollecting each possible context. A familiarity parameter describes the discrimination between each source in the absence of recollection. There is some evidence that familiarity may contribute when the source being recalled is encoded as a property of the cue itself, rather than as an external association (Diana et al., 2008). Alternatively, in a standard source task, for which the correct source is always one of two possible options, a participant could probe and compare the familiarity of each cue-source combination by mentally reconstituting each one. In other circumstances, the familiarity parameter may also reflect the contribution of other (potentially non-mnemonic) evidence. While it is not clear whether

familiarity should be diagnostic in such a task, there is a general expectation that, like associative recognition, performance should be comparatively reliant on recollection and that (relatively) linear ROC curves should therefore be found.

In each of these cases, the DPSD model employs few parameters, but their psychological implications are quite clearly defined. This, perhaps, is one of the primary reasons that the DPSD model has been so widely adopted as a means of quantifying and interpreting behavioural data. In particular, the model has been used to interpret the results of lesion and imaging studies, in an effort to ground the phenomenological descriptions of recollection and familiarity in terms of neurobiology (Aggleton et al., 2005, Cipolotti et al., 2006, Fortin et al., 2004, Sauvage et al., 2008, Yonelinas, 2002a).

Not surprisingly, given its involvement in the interpretation of such studies, the accuracy of the DPSD model has since come under increasing scrutiny. One criticism frequently leveled at the DPSD model is that its description of recollection as thresholded is unrealistic, requiring perfect recollection or none at all (Dunn, 2004, Rotello et al., 2005, Slotnick and Dodson, 2005, Wixted and Stretch, 2004). There is perhaps some legitimate confusion on this point. In the original proposition of the formal DPSD model, recollection is indeed described as 'all-or-none' (Yonelinas, 1994). Nonetheless, even in the same sentence the definition of recollection is somewhat more qualified (p.1343, emphasis ours):

> ... for any item, the subject either succeeds or fails at retrieving *something* about that specific study event.
>
> Yonelinas (1994)

In subsequent papers, Yonelinas and colleagues have emphasised that the term all-or-none is misleading, and that while the DPSD model supposes that recollection can sometimes fail, it is agnostic as to the trial-to-trial strength of successful recollection (Parks and Yonelinas, 2007, Yonelinas, 1997, Yonelinas et al., 2010). Strictly speaking however, this nuanced view is at odds with the formal DPSD model itself. One problem, perhaps, is that the DPSD model was originally designed to account for ROC data with a relatively small number of points. In this case, the assumption that all high strength, recollected trials will be assigned the maximum possible confidence is more likely to be met. In contrast,

finely grained estimates of confidence will dissociate weak from strong recollection more frequently, e.g. (Mickes et al., 2009), particularly when participants are instructed to spread their responses evenly across the scale. Some researchers have attempted to avoid this problem by instructing participants to restrict trials for which recollection occurs to a single confidence rating (e.g. Montaldi et al., 2006). This approach, however, offers little advantage in general over the standard remember-know procedure in terms of isolating familiarity from recollection, and is therefore subject to many of the same disadvantages such as sensitivity to task instructions (see Section 1.4.1).

Proponents of the DPSD model argue that while there are likely to be boundary conditions where the model breaks down, in most recognition tasks the result of any successful recollection will usually lead to high confidence that the stimulus was previously encountered. This assumption may well be approximately true for item recognition tasks, since even weak recollection might provide strong evidence for oldness. It is arguably less likely to be true for associative or source tasks where information may need to be recalled accurately in order to provide diagnostic evidence. Consider, for example, the type of cued recall required to recognise that a pair of items A-C is a rearranged combination of items from the study phase. The participant in this case might recollect that A was originally paired with B and that therefore the pair A-C must be rearranged. This 'recall-to-reject' could plausibly yield varying levels of confidence depending on how similar B and C are or how strongly the original pairing of A-B is recalled, even while the prior occurrence of A in some pairing might be declared with very high confidence. Similarly, in a source retrieval task participants often have to remember whether a cue was originally presented in some context. Successful completion of the task is heavily and directly dependent upon recollection of this information; weak recollection in this task should therefore lead to low rather than high confidence.

In support of the view that recollection (and confidence based on it) can be graded in such circumstances, recent studies have demonstrated curved ROCs for both associative (Mickes et al., 2010) and source tasks (Rotello et al., 2005, Slotnick, 2010, Slotnick and Dodson, 2005). The authors of these papers argue that this curvilinearity is directly related to recollection, not familiarity, and that therefore the DPSD model is flawed in its description of recollection as a

probabilistic variable.

Further evidence comes from dual-process neurocomputational models (Elfman et al., 2008, Greve et al., 2010, Norman and O'Reilly, 2003), which predict variable, though bimodally distributed, recollection strength. In Norman and O'Reilly (2003) the prediction is derived from an assumption that recollection can be described by a pattern-completion of memory representations in sparse hippocampally-inspired networks, though in (Greve et al., 2010) the authors found a similar bimodal pattern of recollection strength using a simple, fully connected Hopfield network. In contrast, in both models the distribution of a proposed familiarity signal was normal and unimodal, and thus it is important to note that while neither model supports the description of recollection as all-or-none, they are entirely compatible with a dual-process view of episodic retrieval.

The debate here may seem anachronistic, since the DPSD model is only intended to provide a useful division of memory into different properties, not a complete and perfect match to empirical data. In other words, so long as the model parameters capture these properties to an adequate extent, arguments about the exact fit are at best unnecessary, and at worst risk rejecting useful scientific data. Here we, like others (Healy et al., 2005, Macho, 2002, Sherman et al., 2003), take a slightly different view. Much confusion has arisen over the fact that the dual-process theory does not equate to the way it is measured: criticisms of the DPSD model are frequently taken as criticisms of the dual-process theory on which it is based, just as the model weaknesses are seen (wrongly) as evidence against the broader theory that recollection is thresholded. From this perspective it is incumbent on proponents of a dual-process theory, in which recollection can fail, to provide a mathematical model which more accurately describes the theory on which it is based.

There is also a second, more practical, reason why the debate should be settled. Given that there is good evidence and broad agreement that recollection can vary in strength (if not the extent of this variation, or how it might relate to confidence in a given task) it makes sense to explicitly measure this variance. This can be done simply by modelling recollected trials as a distribution of memory strength, just like lures or familiar trials. Two different outcomes are possible here. The DPSD parameters might closely match estimates of recollection and familiarity drawn from the more complex model, in which case the DPSD model

may continue to be used to estimate familarity and recollection in a given task, with the advantage that boundary conditions and parameter robustness would be quantified much more clearly than is currently the case. Alternatively, the DPSD model parameter estimates may not reliably match those from the more complex model, indicating that the extent to which they capture recollection or familiarity is subject to error, and that the DPSD model should be updated to address this problem. It is therefore important to consider a set of signal-detection models which amend the DPSD model to this effect.

2.3.3 Mixture signal detection models

The DPSD model can be adjusted by retaining the assumption that two different types of memory response exist, based on recollection and familiarity, but describing each one as providing variable evidence for the given task. The result is a mixture signal detection model (Figure 2.8) in which familiar and recollected trials are each represented by separate gaussian distributions, of which recollection is the stronger. A mixing parameter, λ, determines the proportion of trials which fall into this strong distribution; it is equivalent to $p(R)$ from the DPSD model. The two distributions are separated from the lure distribution by d'_F and d'_R lure standard deviations respectively, and the variance of each is denoted by $v(F)$ and $v(R)$.

We shall refer to this model as the dual-process mixture signal detection (DPMSD) model, for the reason that two separate types of memory (familiarity and recollection) are explicitly assumed. As we shall shortly review, other mixture models may be postulated which do not make this assumption. The labelling of the DPMSD model does not necessarily denote a difference in its mathematical structure compared to the alternative interpretations discussed below, but does specify the psychological meaning of its parameters in this particular way and also results in distinct predictions for the model's behaviour in different memory tasks. The DPMSD model can be extended to account for the pattern of data in associative or source tasks. Much like the DPSD model, this extension involves separating recollection into two possible types: recall-to-accept and recall-to-reject in associative recognition, and recollection of each context in source tasks.

A variable recollection dual process, or VRDP model (Onyper et al., 2010, Sher-

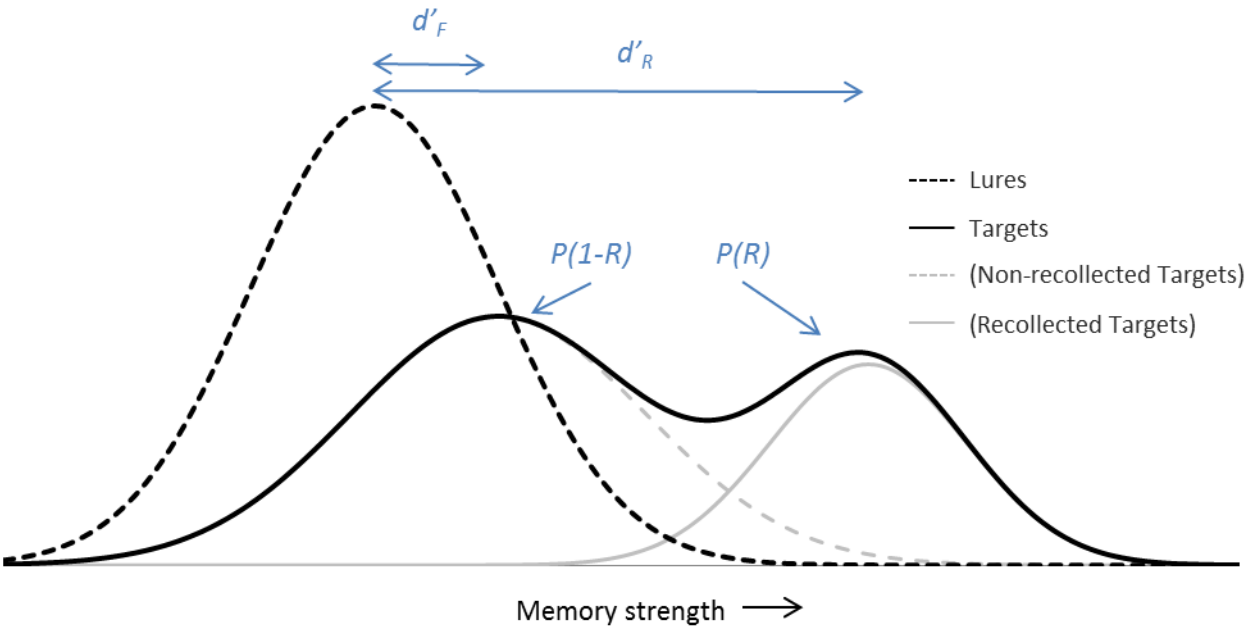


Figure 2.8: The dual-process mixture signal detection (DPMSD) model. Here, as in the DPSD model, recollection is assumed to be thresholded and occurs with probability $p(R)$, but the memory strength for these trials can vary and is normally distributed with mean d'_R and variance $v(R)$. The remaining $p(1-R)$ non-recollected targets are distinguished from lures using other evidence (including familiarity) with mean strength d'_F and variance $v(F)$.

man et al., 2003) is almost identical to the mixture model described above, except that it makes the slightly stronger assumption that the variance of all three distributions are the same. It thus deviates from the standard DPSD model only by a single parameter, the recollection strength d'_R. We argue that this equal-variance assumption should not be made a priori, for two reasons. Firstly, the model is defined on a confidence scale which may not relate linearly to memory strength, and thus variance at high confidence may not be equivalent to the same numerical variance at a lower confidence. Secondly, under a dual-process interpretation, assumptions about the variances of each distribution implicitly require assumptions about the interaction of recollection and familiarity. For example, if evidence from recollection and familiarity are combined when making a decision, the variance of the recollection distribution should be strictly greater than that of the familiarity distribution (since it comprises strictly greater sources of evidence). If, instead, evidence from recollection is weighted more heavily or

used to the exclusion of familiarity, as a DPSD model predicts, the recollection distribution may have a lower variance than the familiarity distribution.

2.3.4 Interpretations of mixture signal detection models

A mixture model implies a thresholded pattern of responses, that is, responses can be divided into two or more distinct types. It similarly implies a graded pattern, since each type of response can give rise to memory or confidence of variable strength. The most immediate conclusion one might draw from the thresholded and graded pattern of responses typified by the mixture model is that the strong memory process, i.e. recollection in the DPMSD model, is correspondingly both thresholded and graded. In actual fact the theoretical significance of both properties remains a matter of some debate, and so mathematically identical mixture models can be distinguished both in terms of their different parameter interpretations, and predictions they make regarding memory. In particular, this class of mixture models, which are both thresholded and graded, do not in themselves require that either the threshold or variability are determined by processes at retrieval. Either characteristic could equally be introduced at encoding, retention or indeed be unrelated to memory processes entirely.

Indeed, while there is some emerging agreement that distributions of confidence or memory strength for previously-seen stimuli may often be well described by types of mixture model (DeCarlo, 2002; 2003, Hautus et al., 2008, Kelley and Wixted, 2001, Mickes et al., 2010, Onyper et al., 2010), in most of these cases the authors concluded that recollection was actually a continuous process, i.e. one which does not fail at retrieval. By one such explanation (DeCarlo, 2003, Hautus et al., 2008), the threshold simply reflects a dichotomy at encoding in that some targets are not attended to during study, consequently no information about them is encoded and they are essentially no different to lures at test. Thus, although a threshold may be observed at retrieval, it cannot be properly described as a memory threshold, nor be isolated to a particular memory process such as recollection.

This version of the mixture model, if accurate, is not only consistent with a continuous account of recollection, but is also entirely compatible with the single-process signal detection view of episodic memory. In this case memory strength in

general is a continuous, graded variable which always provides some information for a previously-attended stimulus. Such a single-process account does, of course, also rely heavily on some untested assumptions, the most important of which is that participants in these experiments must have failed to attend to a substantial proportion of trials at study: commonly between 25% and 50% of trials in the experiments cited above. An important implication is that any memory threshold reported in the literature must also have resulted from poor attention at study (or a similar non-mnemonic artefact of the data), since memory strength itself is continuous and does not readily explain a thresholded pattern.

Furthermore, a single-process interpretation results in the strong prediction that the lure distribution and the lower strength target distribution should be matched in strength, or at least that their separation, measured by d'_F in the full DPMSD model, should be very close to zero and reflect only non-mnemonic evidence. This prediction is incompatible with the results from Onyper et al. (2010), however, which directly compared this model in item recognition with the VRDP model, and found that lower-strength targets had significantly greater strength than lures. In other words, this study found evidence of a second, weaker source of evidence (presumed to be familiarity) which supported item recognition even when recollection failed.

Given the difficulties of reconciling the data to a single-process account, some researchers appeal to a related yet subtly different version of this argument, allowing that a memory threshold might exist, but that it is entirely determined at encoding (Kelley and Wixted, 2001, Mickes et al., 2010). Here, even items that are attended to may not be sufficiently encoded for later memory to be successful. Furthermore, by a dual-process perspective this 'thresholded encoding' may be isolated to a particular process. For example, minimal perception and attention might be sufficient for some level of familiarity to later distinguish studied from non-studied items, whereas later recollection might only be possible following encoding for a particular - greater - length of time.

It is worth noting that the thresholded encoding explanation still requires encoding failure to occur for a large proportion of trials at study, and is also to some extent at odds with the variable encoding explanation used by the same authors to justify the UVSD model, which specifically describes encoding as graded (Wixted, 2007a). Unlike the attention explanation, however, this view is con-

sistent both with dual-process theory and the results from Onyper et al. (2010), since it allows that two separate, and diagnostic, sources of evidence may contribute to item recognition. In fact, the 'encoding threshold' model differs from the DPMSD model primarily in terms of interpretation, rather than its mathematical structure[3]. The most fundamental difference is that it implies that recollection is thresholded at encoding rather than at retrieval. In other words, recollection is thresholded only in the sense that relational information may not be encoded during a study phase, and is therefore unavailable for some trials at test (Kelley and Wixted, 2001).

Finally, an alternative class of model (which can, like the UVSD and mixture models, be interpreted as either single or dual-process) can mimic the pattern predicted by a mixture model, by introducing a skew to the target distribution. One particular example of this is the hierarchical Relational Binding Theory (hRBT) model (Shimamura and Wickens, 2009), in which the skew is theoretically motivated by a nonlinear increase in memory strength. By this account, a small number of strong memories tend to be far stronger than the more common weak memories, resulting in a skew towards the high strength end of the evidence scale (modelled by an ex-Gaussian distribution). While the hRBT model, and skewed target distribution models in general, predict similar target distributions to those predicted by the mixture model, there are some differences. We examine these, and assess which model our data supports, in the general discussion (Chapter 10).

Since the earliest use of signal-detection theory as a framework for modelling the relationship between memory signals and behaviour, the simple equal variance model has given way to a plethora of competing accounts. Some of the most influential of these have been outlined in this Chapter, focusing on differences between them; nonetheless it is also important to highlight the areas of agreement between models. Firstly, none of these models have deviated from the original

[3]A possible exception is for associative recognition tasks, where rearranged pairs depend on two study episodes, but intact pairs only one. If some proportion λ of study episodes were encoded, then λ^2 rearranged test trials will have access to both episodes, $(1-\lambda)^2$ will have access to neither, and the remaining trials will have access to one. For intact pairs, λ will have access to the original study presentation, and $1-\lambda$ will have no access. Thus rearranged pairs should follow a mixture of three Gaussian distributions, and intact pairs a mixture of two. To our knowledge, no model of associative recognition has incorporated this prediction of thresholded encoding, including (Mickes et al., 2010) in which thresholded encoding was explicitly preferred as an explanation for the results.

assumption of the EVSD model that memory strength is normally distributed by default. Secondly, in item recognition at least, there is strong and consistent evidence that old items are associated with a greater range of confidence than new items, and each of the types of model, DPSD, UVSD, hRBT and MSD, incorporates at least one parameter which achieves this difference. Furthermore, each model has an interpretation which ascribes this difference in variance to the influence of multiple distinct types of evidence. Although the UVSD and certain MSD models can also be interpreted in terms of a single process, in both cases there is empirical evidence against that specific interpretation, as well as considerable extant evidence from other fields for the existence of separable recollection and familiarity signals (see for example Yonelinas, 2002a for a review).

Thus the focus of the argument, at least for those who broadly agree that separable sources of evidence contribute to recognition, has moved on to the question of how to characterise the stronger of these sources. In particular, does recollection have a threshold (at retrieval) or is it continuous? Just as importantly, and separable from the question of whether or not it has a threshold, is successful recollection graded?

2.3.5 The central question: characterising recollection

It is important to determine whether recollection is thresholded or continuous, and whether successful recollection is graded. If recollection has a reliable, measurable threshold, and familiarity does not, then behavioural ROC data can be used to separately estimate the contributions of each process. If instead recollection is continuous then quantitative measures derived in this way, which are used to interpret findings from imaging, lesion and animal studies in cognitive terms, would be invalid. Whether recollection is graded or not is also important, since it determines whether the strength and rate of recollection may be dissociated and separately examined, or whether a single parameter should capture most of the observable effect of recollection on behaviour. As we shall outline in this section, however, determining these properties requires careful adherence to three principles: separation of the strength and frequency of recollection; direct examination of accuracy or memory strength; and separation of the effects of encoding and retrieval.

In many cases, the arguments for each of the properties of recollection that determine how it should be modelled - a threshold and graded strength - have been conflated. One recent example is a paper by Slotnick (2010), in which the author examined source ROCs in three experiments, and concluded that they supported a continuous model of recollection, albeit one in which source misattribution played a significant role. In each experiment, however, the analyses were restricted to those trials for which there was already evidence of successful recollection, i.e. those for which participants had responded 'remember' on a remember-know-guess judgment or, in the third experiment, those with the highest possible old/new confidence rating[4]. Thus the three nonlinear source ROCs produced are compelling evidence that successful recollection can vary in strength, but they cannot provide any evidence for or against a threshold in general, since by definition only above-threshold trials were examined. It is important, if progress is to be made in characterising recollection, to explicitly separate the strength of recollection (which may be graded) from its frequency (which may reflect either a thresholded or continuous process).

The assumption that recollection is graded has been less frequently challenged in the recent recognition literature, but some researchers have argued that the main source of evidence for this assumption - curvilinear ROC curves - do not require a graded underlying process (Bröder and Schütz, 2009, Malmberg, 2002). By this view, recognition (for example) might be explained in terms of finite discrete states, with different confidence ratings simply comprising different proportions of each state rather than different values along a continuum. In this case, it is argued, the curvilinear ROCs which have provided such compelling evidence for signal detection models over high-threshold alternatives may in fact be compatible with threshold models after all. We have some sympathy with aspects of this view, which highlight problems with the assumption that confidence ratings provide a direct reflection of memory strength.

As a consequence, high threshold (i.e. non-graded) models may provide a valid account of the extant behavioural data (Atkinson and Juola, 1974). The demonstration of graded accuracy across finely grained confidence values might be more

[4]Helpfully, however, the paper does include the raw data required for the reader to be able to form an averaged source ROC for all of the responses in each experiment. In fact, when the full dataset is analysed the results are at least compatible with a mixture model description of recollection, and arguably support it more than they do the author's preferred UVSD model.

challenging for this account (Hilford et al., 2002, Mickes et al., 2009), requiring a relatively high number of parameters to explain the mapping from discrete memory to graded responses. More broadly, however, is it doubtful that variability in confidence alone can provide unequivocal evidence of graded memory strength, since it might also or instead be a product of assigning it a rating, as illustrated in Figure 2.9. One illuminating way to consider the issue is that interpreting graded confidence as reflecting only graded memory strength (i.e. not metacognitive noise) requires that each participant uses the confidence scale so consistently that all responses assigned a confidence N reflect stronger memory than every response $< N$, and weaker memory than every response $> N$, an assumption we believe is highly unlikely to hold in practice. Thus, the question of whether memory strength is itself graded should be tested by examining it more directly, for example in terms of its accuracy instead of perceived strength or confidence.

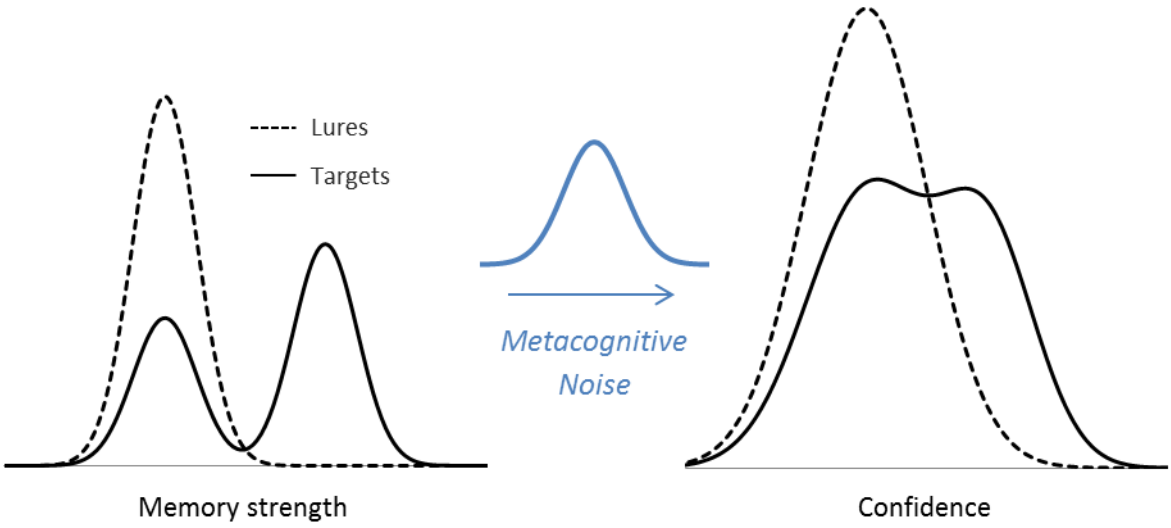


Figure 2.9: Effect of gaussian noise from a metacognitive assessment of memory, e.g. assigning confidence ratings. As we illustrate here, such metacognitive noise can make a thresholded pattern of memory strength appear continuous - even when the original distributions are well separated. Confidence is not perfectly correlated with memory strength; it depends on the task as well as non-mnemonic factors such as preceding responses, mood, hand position, motor error, motivation, attention, or subtle changes in rating strategy across an experiment, inevitably introducing noise.

The question of whether or not recollection is thresholded also goes beyond the

existence (or not) of a thresholded pattern in memory strength or confidence data. As we have noted, this pattern could also be interpreted as reflecting a threshold at encoding, which could be either mnemonic (Kelley and Wixted, 2001) or non-mnemonic (DeCarlo, 2003). In contrast, neurocomputational models (Greve et al., 2010, Norman and O'Reilly, 2003), as well as several dual-process models (Diana et al., 2007, Eichenbaum et al., 2007, Montaldi and Mayes, 2010) describe recollection as being thresholded in terms of retrieval, in that it can fail regardless of the presence or absence of information. To our knowledge, a test that directly differentiates between these competing interpretations has not been reported, despite the importance of the point in distinguishing between models. It is in fact relatively straightforward to test, by comparing recollection across two conditions which are matched in terms of encoding, but which differ in terms of retrieval. If a threshold is determined at encoding, the probability of recollection should remain constant across two such conditions. If instead a threshold exists at retrieval, probabilities of recollection should be affected by the manipulation and differ across the two conditions. Precisely the same logic can be used to determine whether recollection provides variable information because the retrieval process is graded, or simply because variable information is encoded at study and therefore variable information is later (discretely) retrieved.

In Chapter 4 of this thesis, we characterise recollection by following the three principles laid out above: We separate the rate of recollection from its strength; we examine the accuracy of recollection rather than participant confidence and we explicitly test whether the rate or strength of recollection are determined at encoding.

Chapter 3

General Methods

This chapter can be divided into two sections. In the first section, Experimental Procedure, we outline the procedures, stimuli and other experimental details which are used throughout this thesis. In the second section, we describe and justify the ROC fitting approach taken in Chapters 5–9.

3.1 Experimental Procedure

Here we outline the general procedure followed for each experiment, as well as those procedures specific to Chapter 4 and Chapters 5–9 separately. All of the experiments in this thesis were implemented using E-Prime (Psychology Software Tools; www.pstnet.com) and displayed on a standard desktop monitor. All experiment screens had a resolution of 640×480 pixels.

Participants were all right-handed volunteers between the ages of 17-31, with no known neurological problems and normal or corrected-to-normal vision. All participants were native English speakers, and those participating in experiments using Christian names as stimuli (Chapters 5–9) were also born and resident in the UK or ROI. For the first hour of each experiment, participants chose to receive either Psychology course credits (up to 2 full credits in each case) or £5. Time beyond the first hour was compensated at a rate of £5 per hour.

All experiments were approved by the University of Stirling Department of Psychology Ethics Committee, and participants gave written informed consent before

beginning the experiment. Within each experiment, participants were given identical written instructions, in addition to a verbal explanation of the procedure. To further ensure consistency across subjects the same experimenter (i.e. the author) carried out the experiment for every participant. Participants were fully debriefed and answers were provided to any additional questions they had about the experiment and its aims.

All model fits in this thesis were performed in Microsoft Excel using the GRG Nonlinear method from the Solver package (Frontline Systems, Inc; www.solver.com) and checked using the fminunc function in MATLAB (The Mathworks, Inc; www.mathworks.com) which gave comparable results. We used maximum likelihood estimation to find parameter values for each model and each dataset; in practice we minimised the negative log likelihood of the observed data under the assumptions of each model.

The nature of the maximum likelihood procedure disallows a guarantee that the global maximum likelihood (or equivalently, minimum negative likelihood) was found for every combination of model and data; however we used multiple starting points to mitigate the chances of accepting local, non-global, maxima. The model fitting approach used is discussed in more detail in the second part of this chapter; in the remainder of this section we outline some aspects of the experimental procedures carried out in Chapters 4 and 5–9.

3.1.1 Chapter 4 procedure

For Chapter 4 responses were gathered using a standard computer mouse and a keyboard was used to advance screens during instructions and between trials. Instructions and lexical stimuli were presented in bold black 18-point Courier New typeface against a white background (Chapter 4).

The full set of cue stimuli consisted of 324 words selected from the MRC Psycholinguistic Database (www.psych.rl.ac.uk, Coltheart, 1981). The words were selected to be of moderate and similar length (5-7 letters) and relatively low concreteness and imageability; this was to encourage reliance on explicit recollection and relational memory for the source retrieval task by minimising perceptual differences between the words and discouraging participants from visualising each as an object in the paired location. Experiment 1 used a reduced set of 232 words

with similar properties; full statistics for each dataset are included in Table 3.1 below.

	Len	Con	Ima	Fam	K-F	T-L
Full set	6.2 (0.7)	371 (85)	419 (62)	489 (69)	26 (30)	202 (229)
Reduced set	6.3 (0.7)	350 (63)	414 (55)	512 (51)	34 (31)	267 (240)

Table 3.1: Mean statistics for the word sets used in Chapter 4, with standard deviations provided in brackets. The full set consisted of 324 words, and the reduced set consisted of 232 words from the full set. Statistics were derived from the MRC Psycholinguistic Database, Coltheart (1981). Len = word length; Con = concreteness rating (100-700, database mean = 438); Ima = imagability rating (100-700, mean = 450); Fam = printed familiarity rating (100-700, mean = 488); K-F = Kucera-Francis written frequency; T-L = Thorndike-Lorge written frequency.

Target stimuli consisted of locations on a grey circle outline of radius 200 pixels, marked by a black cross. Locations were defined by their angle in integer degrees, making 360 possible targets around the circle. Of these 360 possible angles, those relating to identifiable points on the circle (e.g. multiples of 45°) were removed, and then other targets were removed to ensure an approximately uniform distribution around the circle for the full (324) and reduced (232) sets of stimuli. The mean distance between two directly adjacent targets (i.e. the distance corresponding to 1°) was 3.5 pixels. In practice, within each study/test block each a minimum distance of 10° (35 pixels along the circle arc) was maintained between each pair of targets.

Responses from participants (i.e. screen co-ordinates selected using the mouse) were first converted into an angle in degrees from the centre of the circle. To maximise precision in the error statistic, this angle was compared to the target angle (itself recalculated to reflect the actual angle of the pixel on which the target cross was centred) to provide the error statistic in degrees. Some aspects of the procedure varied slightly across the two experiments, and it is therefore outlined in detail in Chapter 4.

3.1.2 Chapters 5–9 procedure

For Chapters 5–9 all responses were collected using a 5-button Psychology Software Tools Serial Response Box. Instructions and lexical stimuli were presented in bold white 18-point Courier New typeface against a black background.

Stimuli consisted of equal numbers of Christian names and abstract images. Christian names were selected on the basis of being readily identifiable as a real name, i.e. adult participants from the UK or Ireland should be to some extent familiar with them. Thus, we included names given to large numbers of US or UK children born between 1950 and 1990 (see for example www.ssa.gov). We included both given names and common shortenings. Where a given name and a shortened version were clearly discriminable from each other we included both forms (e.g. Tony and Anthony were both included), where two names or forms were similar we did not include both (e.g. Diana was included, but Diane was not). Names were screened for length, either 4-8 letters long in experiments using the larger sets of 480 (Chapter 9) or 432 names (Chapters 6 and 7), or 4-7 letters in the experiment using the reduced set of 324 names (Chapter 5).

Abstract images were derived from textures and photographs downloaded from image*after (www.imageafter.com), a royalty free online image bank. A total of 575 images were selected, manipulated and cropped to 50×80 pixels to remove identifying features and minimise their semantic content. These 575 were then rated for abstractness by 9 participants (4 female; mean age 24.3, range 19-31) and the 324/432/480 most abstract were selected for the experiments in Chapters 5–9. The results of these ratings are summarised in Figure 3.1, which also shows the cut-offs for both the full (480) and reduced (432/324) sets of images used.

Names and images were used to form stimulus pairs in three configurations (Figure 3.2): name-name pairs, image-image pairs and mixed name-image pairs. For name-image pairs, images and names each appeared in the top half of the screen 50% of the time. All stimulus presentations were preceded by a fixation cross and were presented in the centre of the screen. At a viewing distance of approximately 1 metre the items in each stimulus pair together subtended a maximum visual angle of 3.7° vertically and 3.4° horizontally.

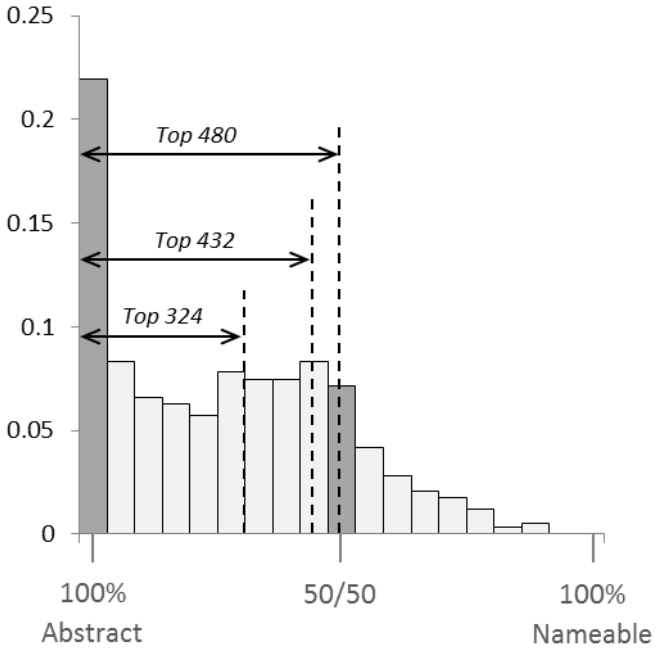


Figure 3.1: Mean rating distribution for abstract images. Nine participants were shown each image in random order and rated it on a three point scale: 'abstract', meaning no part of the image was nameable or identifiable; 'somewhat nameable', meaning they felt that with time they would be able to assign a name or label to the image; or 'nameable', meaning they could readily identify or assign a name to part of the image. The full set of 480 images used in Chapter 9 were rated 'abstract' 70% of the time; the 432 images used in Chapters 6 and 7 were rated 'abstract' 74% of the time; and the 324 images from Chapter 5 received 'abstract' ratings 93% of the time. Filled bars indicate images with maximum, exactly average or minimum abstractness ratings.

Figure 3.2: Example stimuli for Chapters 5–9. Names and images were studied in pairs in the centre of the screen. Each pair consisted of either two names, two images or one name and one image, as shown. The vertical order of names and images in these final mixed pairs was reversed for 50% of presentations.

3.1.3 Measuring performance using discrimination

Discrimination, for an equal variance signal detection model, is defined as the separation of target and lure distributions in standard deviation units and is denoted by d' (see Section 2.2.2). Discrimination can be calculated empirically in at least two distinct ways: from the confidence or strength data (either directly or by fitting the data to a measurement model such as EVSD) or from the y-intercept of the (linear) z-ROC.

The standard discrimination statistic, d', is predicated on the assumption that the target and lure distributions are well described by an EVSD model, i.e. that both distributions are normal and have equal variance. Given that there is considerable empirical evidence that memory strength is more variable for targets than for lures, a more accurate measure of discrimination under these circumstances may be given by d_a, which describes the separation in terms of combined lure and target standard deviations Macmillan and Creelman (2005). Specifically, when targets have greater variance than lures, a greater proportion of targets will overlap with the lure distribution and d' will overestimate their separation.

We use d_a as a standard bias-invariant measure of performance throughout this thesis for this reason. Where μ_T and μ_L denote the mean confidence rating to intact and rearranged pairs respectively, and σ_T and σ_L denote their standard deviations, d_a is calculated by:

$$d_a = \frac{\mu_T - \mu_L}{\sqrt{(\sigma_T^2 + \sigma_L^2)/2}}$$

3.2 Fitting and comparing measurement models

Many aspects of the approach taken to model fitting can have a profound effect on the pattern of fits and, as a result, the conclusions drawn. In this section we detail and justify the strategy used in this thesis to fit confidence data to the models outlined in Chapter 2.

3.2.1 Fitting independent and individual data

Firstly, we fit all empirical data before collating it to form ROC curves. As noted in Section 2.2.5 points on the ROC curve are not independent of each other, thus fitting directly to underlying confidence ratings is a more appropriate approach than fitting to the ROC or z-ROC.

Secondly, it is important to recognise that the models detailed in the previous chapter describe memory performance for an individual; collapsing data across multiple participants and fitting the models to this group data can introduce averaging artefacts (Brown and Heathcote, 2003, Wickens, 2002). One important problem is that the confidence scale is relative, not absolute. In other words, the relative relationship that 8 denotes a higher confidence than 7 holds for each participants set of responses; in contrast the absolute memory strength value of a given confidence rating is not fixed, it is different for each participant. It is therefore unjustified to equate confidence ratings across participants, as happens in the construction of a group ROC. A second problem, which is particularly critical if thresholded and continuous models are being compared, is that thresholded data averaged together may appear continuous if the thresholds are not identical. Thus all ROC data are fit at the level of individual subjects, not at a group level.

3.2.2 Criteria

Since each model was fit at the level of individual subjects, the position of criteria (which determine the relationship between latent memory strength and confidence response) were allowed to differ between participants. We fixed criteria within participants, however, across conditions. Using the same criteria values for every condition performed by a participant is equivalent to assuming that they use the confidence scale in a consistent way across these conditions. Note that the correspondence between underlying memory strength and confidence is not required to be constant across all trials, a much stronger assumption. Since reported confidence is likely to be affected by some non-mnemonic factors, including but not limited to metacognitive error, preceding responses or hand position, two trials of the same memory strength can certainly result in different reported confidence values (from the same participant). This has the effect of adding noise to the true

underlying confidence, increasing the variance of each distribution when they are estimated from confidence data compared to a less subjective measure of memory strength, such as accuracy (see Section 2.3.5). Here instead the relationship between memory strength and mean confidence response was assumed not to vary systematically across conditions, i.e. confidence ratings of 6 in condition A indicates trials of similar strength, on average, to those assigned a rating of 6 in condition B.

3.2.3 Model selection

Two methods of model selection commonly employed - hierarchical likelihood ratio testing (hLRT) and Akaike/Bayesian Information Criteria (AIC/BIC) - rely on some particular assumptions about the data and the underlying model which are generally broken when comparing mathematical memory models. Firstly, they each assume that one of the models being tested is correct, i.e. that it gave rise to the data. This assumption is necessarily broken since all models being tested are inevitably (and deliberately) simplifications. Therefore at the very least one should always bear in mind that the results of a model selection can never justify the acceptance of one particular model, but at most its superiority within the subset of models being tested. At worst, breaking this assumption might render the statistics themselves less reliable than would otherwise be the case: for example, larger differences in the fit statistic might be expected when both compared models are incorrect than when one is correct. Some fit statistics make additional assumptions which are unlikely to be satisfied in ROC data. For example, the BIC assumes that the data is fully explained by exactly the number of parameters in the model, resulting in a particularly severe penalty for the introduction of additional parameters, while the AIC allows that additional, less influential parameters exist: a scenario which is arguably more likely to reflect psychological data (Yang, 2005).

A second assumption may be more important. In both cases, the selection approach compares goodness-of-fit (in this case the maximum likelihood of the data across the parameter space) with the number of free parameters in the model; a preferred model will tend to explain the data well with fewer parameters. The reason for this is to avoid over fitting, since a more flexible model may capture

more of the noise in the data, which might then be wrongly taken as evidence for the model as a description of the underlying signal. Importantly, both hLRT and AIC/BIC penalise the addition of every parameter equally. In other words, the addition of a single parameter to the model is assumed to increase its flexibility (and therefore fit) by some amount which is independent of the range of values that parameter can take, or its relation to the other parameters in the model.

In reality, bounding a parameter (for example forcing it to take a positive value) will reduce opportunities for over fitting and therefore limit the flexibility of the model compared to when the parameter can vary freely (Self and Liang, 1987). In addition to this, different parameters will increase the range by different amounts depending on how they interact with existing parameters. Therefore hLRT, AIC, BIC or other methods that assume a fixed improvement in likelihood for each additional parameter can at best be viewed as a guide to model selection.

Given the caveats above, a more accurate guide to model selection might be obtained by directly estimating the flexibility of the models under consideration, using a model recovery simulation, rather than relying on each of the assumptions underlying BIC and AIC being correct. For example, a relatively flexible model with fewer parameters might be preferred using BIC, even when it is not in fact the model generating the data. To gain some insight into how fit statistics might behave when comparing mixture (thresholded) and continuous models, we simulated confidence data from a mixture model and then fit the resulting data to both the original mixture model and the simpler UVSD model. The results of this analysis are summarised in Appendix B, but the important point is that item recognition data generated using a mixture model was more parsimoniously fit by the UVSD model, according to both AIC and BIC fit statistics.

The fact that AIC and BIC both select the simpler UVSD model ahead of the more complex DPMSD model which generated the data, highlights that the most parsimonious description of the data (in this case the UVSD model) and the true underlying model (DPMSD) do not necessarily overlap. While fit statistics can be used as a *guide* to model selection for a given dataset, a better strategy for arbitrating between two memory models would be to test the ability of each to explain performance across a variety of memory tasks and experiments. This is because two parameters which are closely correlated in a single task (thus accounting for the data less 'efficiently' and deemed unparsimonious or unnecessary

by AIC/BIC) might both be required to explain patterns of performance across other tasks.

The analysis here highlights the danger of using confidence ratings in particular to arbitrate between thresholded and continuous models of recognition (as opposed to estimating parameters from a known model, for which confidence data of this resolution may be appropriate). We shall therefore take a different approach in Chapter 4, by testing the critical point of contention between the models (that recollection is either continuous or it is subject to a threshold) more directly, using accuracy data. Before reporting the results in that first data chapter, however, we first introduce the ERP imaging methods which we later employ in Chapter 9.

Chapter 4

Characterising Recollection using a Novel Graded Source Task

Does recollection sometimes fail completely? Intuitively, the answer seems to be yes. The common experience of forgetting where you left your keys fits well with the threshold theories of recollection introduced in Chapter 2 that state retrieval can indeed fail (Sherman et al., 2003, Yonelinas, 1994). Alternatively, however, recollection may be a continuous signal that always provides some mnemonic information; by this view retrieval 'failure' reflects weak, not absent, recollection (Slotnick and Dodson, 2005, Wixted, 2007a). Characterizing recollection correctly is necessary for it to be accurately dissociated from other memory processes. For example, some dual process theories distinguish thresholded recollection from continuous familiarity (Onyper et al., 2010, Yonelinas, 1994), which provides an acontextual sense of oldness. This functional dissociation only remains valid, however, if recollection can be definitively shown to be thresholded. If, instead, recollection is continuous, many existing conclusions within memory research, from the specific decline of recollection in aging (Howard et al., 2006) to the mapping between cognitive processes and neurobiological structures (Eichenbaum et al., 2010, Peters et al., 2009, Ranganath, 2010), may need to be reinterpreted (Wixted, 2007a). Thus, whilst apparently obvious, the fundamental nature of recollection remains highly disputed.

4.1 Introduction

To date, the most important evidence for a retrieval threshold comes from Receiver-Operating Characteristic (ROC) curves (Yonelinas and Parks, 2007). Theoretically, the shape of this curve provides information about the memory processes supporting the task. Classically, a strictly thresholded process should result in a linear ROC curve, while nonlinear ROCs are assumed to result from a continuous process. Linear ROC curves have previously been found in source and associative tasks for which recollection is believed to be critical (Rotello and Heit, 2000, Yonelinas, 1999, Yonelinas et al., 1999), but not those such as item recognition where familiarity should play a stronger role (for a review see Yonelinas and Parks (2007)). This qualitative difference in ROC curves across tasks has been interpreted as strong evidence for the view that familiarity is a continuous process, but recollection is subject to a threshold. More recently however, the same technique has provided evidence in apparent opposition to this view. In a source memory task participants must retrieve some specific context of an episodic memory, such as the side of the screen that a word was presented on. Source tasks should rely heavily if not exclusively on recollection, yet the discovery of nonlinear source ROC curves (Rotello et al., 2005, Slotnick, 2010, Slotnick and Dodson, 2005) has led some to reject the thresholded view of recollection in favour of a continuous model (Wixted, 2007a).

One reason why ROC data has produced apparently incompatible results, and therefore failed to resolve the debate over how to characterize recollection, is that it provides an indirect measure of memory strength. Typically, the different sensitivities required to sample the ROC curve are derived from subjective judgments of confidence by participants. The validity of this approach depends on the assumption that confidence and memory strength are directly and consistently related, a claim which has been questioned (Bröder and Schütz, 2009, Malmberg, 2002). In practice, confidence on a particular trial is also influenced by non-mnemonic factors such as mood, task instructions or preceding responses. The effect of this is to add Gaussian noise to both target and lure distributions, making the ROC data nonlinear and thus apparently incompatible with a threshold account - even if the recollection process supporting it is fundamentally thresholded (Figure 2.9). The factors determining the shape of an ROC curve

are therefore more complex than commonly acknowledged, making it difficult to draw unambiguous conclusions about the nature of recollection from confidence judgments alone.

4.1.1 An alternative approach: Measuring source accuracy

Here we introduce a new approach to measuring recollection, avoiding the subjectivity associated with confidence judgments and relying instead on a more direct measure: response accuracy. We also simultaneously gather confidence ratings; this will allow us both to test how our conclusions would be affected by relying on confidence instead of accuracy, and to directly examine the relationship between the two measures. To assess accuracy we employ a source task that is designed to provide a fine-grained assessment of retrieval error (Figure 4.1). At study a series of words were presented visually, each paired with a unique source location that participants had to reproduce (Figure 4.1(a)). At test, the studied words were re-presented and participants were asked to recollect the associated source location as accurately as possible (Figure 4.1(b)). Thus instead of relying on (subjective) confidence to indicate memory strength on each trial, we use the (objective) error: the arc length in degrees between the recollected and correct source locations (Figure 4.1(c)).

We used locations on a circle for several reasons. Firstly, the experiment is analogous to other source tasks which employ a location as the source which must be recollected (e.g. which side of the screen was the item originally presented). Secondly, the location is an external context for the item, which is important given evidence that internal context (such as a colour) might be retrievable using familiarity (Diana et al., 2008). We further avoided the possibility of unitization by presenting the location and word on separate screens during the study phase (Figure 4.1(a)). Thirdly, the location is unidimensional, in the sense that it is difficult to separate it into distinct components which might be recollected separately from each other. This is important because for complex stimuli where this is possible, such as an image or a word, thresholded recollection of different parts of the stimulus could give rise to more continuous-like overall strength. For example the first and last letter of a word (but no other letters) might be recollected, producing overall memory strength somewhere between complete failure

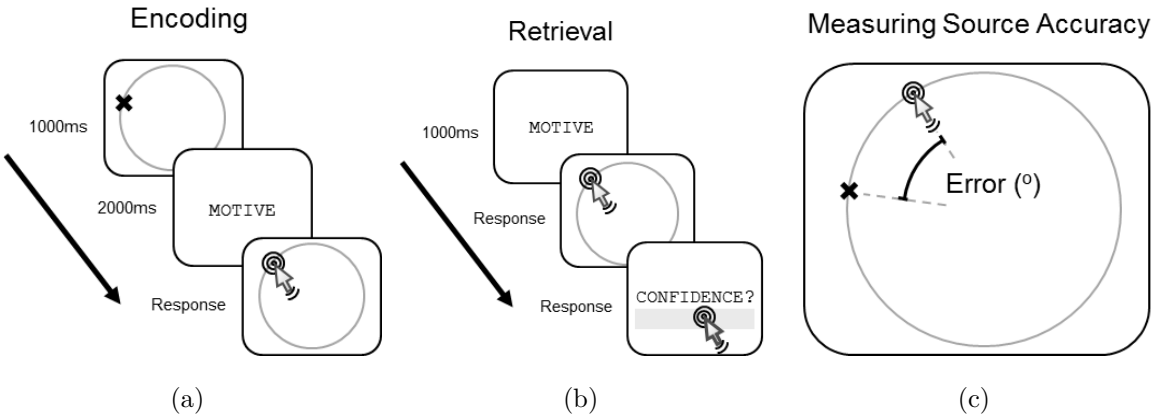


Figure 4.1: Novel source memory task. Study trials, (a), required participants to memorise unique word/location pairs, indicating the location after each trial to confirm attention and provide a baseline measure of response error. Test trials, (b) required participants to indicate the recollected source (location) for each studied word, and rate their confidence, using a mouse. Response errors, (c), were measured by calculating the arc length in degrees between the correct and recollected locations.

and complete success. Finally, the use of a circle makes all of the possible locations equivalent in terms of their relationship to each other, meaning that chance performance is clearly defined as 90° from the target on average (the definition becomes much more complex for, e.g., Euclidean space where participants could influence their chances of being near the target by selecting a more or less central location).

4.1.2 Developing a model of source accuracy

In order to detect a threshold, we first define the expected results of the null hypothesis; in this case that recollection is a continuous process which provides some variable memory strength on every trial. The overwhelming majority of continuous signal-detection memory models define this memory strength to be approximately normally distributed (Figure 4.2(a))[1]. What does this simple continuous model predict about the error distributions we should expect? Strength should be inversely related to mean error, insofar as higher strength memories

[1]Here we assume normally distributed memory strength, but results did not qualitatively differ when we used a Weibull distribution instead, as some have suggested might be a better description of memory strength (Wixted, 2007a).

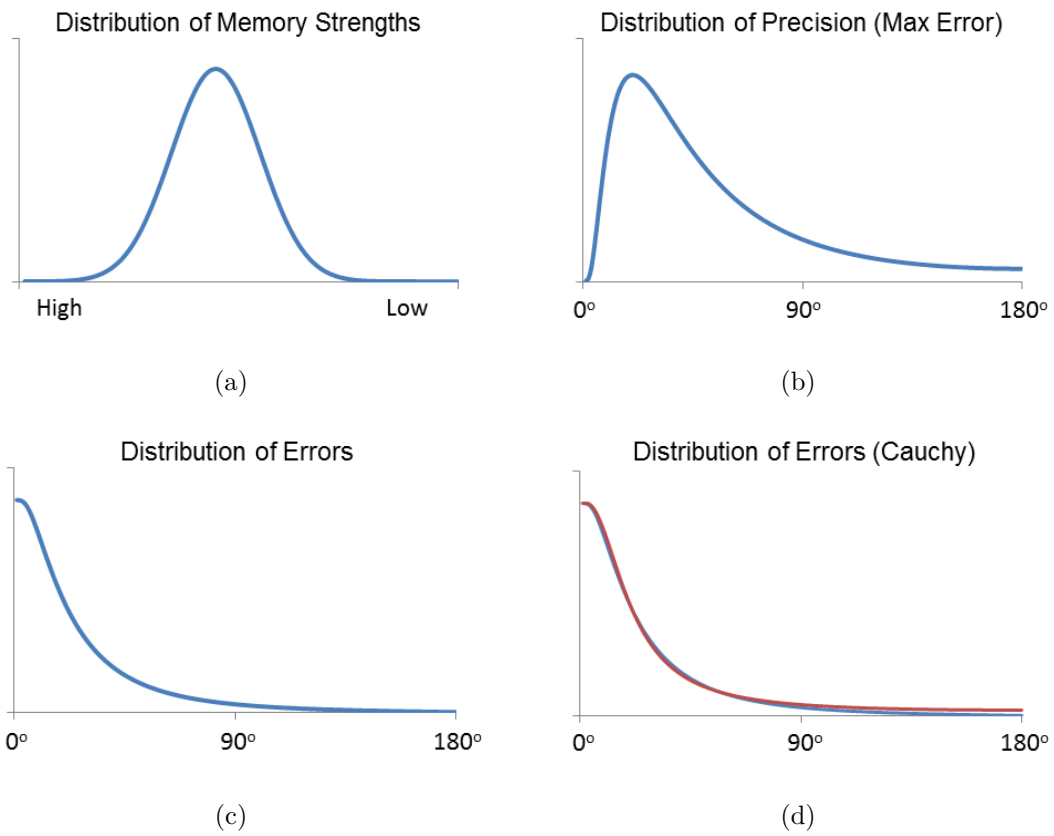


Figure 4.2: Expected error distribution arising from Gaussian memory strength. Assuming memory strength is approximately normally distributed, (a), and that increasing precision requires exponentially stronger memory, maximum errors (precision) on each trial should be lognormally distributed, (b). This leads to an expected error distribution as shown in (c), which closely approximates a Cauchy distribution, overlaid in red in (d).

should be associated with lower error on average. To outline this relationship clearly we relate both properties to an intermediary, theoretical concept of precision, defined by the maximum error possible on an individual trial. One important point to note here is that error and precision are not simply related on a trial-to-trial basis. In particular, a very small error on a trial does not require that the memory strength was high, since a low strength, imprecise trial could result in an accurate response through chance alone. The simplest example is the complete absence of memory; responses will be randomly located relative to the target and may therefore land any distance from 0° to 180° away from the target. This is analogous to the understanding that hits in a simple yes/no recognition task represent both successful memory and correct guessing, and should be corrected

for by examining false alarms.

Precision can however be related directly to mean error. Suppose a location is known to within $\pm n$ degrees. Then assuming the target location lies within this range, the maximum absolute error will be $n°$ and the minimum will be $0°$. Since precision is used here as a theoretical concept, we can simplify it by assuming that the target location has a uniform probability distribution within the range of known values. Although it may seem more intuitively reasonable that the probability distribution should be graded and not uniform, by modelling it as graded here we would end up having to measure the variability of memory strength in two places: the memory strength distribution, and our (arbitrary) relationship between error and precision. By instead treating it as uniform we move all of the 'gradedness' of each memory to the strength distribution, where we can measure it explicitly. By simplifying the model in this way precision is straightforwardly defined as the maximum error for each trial and is linearly related to mean error: for a trial known to within $\pm n$ degrees, the mean absolute error will be $n/2$ degrees.

We can therefore relate memory strength directly to mean error. High memory strength can be regarded for this task as having a more precise knowledge of the true location. For example, a participant might recall a location to within a few degrees, or only remember the approximate quadrant of the circle it appeared in. For the purposes of these analyses we assume that the first, more precise instance reflects greater memory strength than the second (though see Section 2.2.1 for a brief discussion of the limitations of memory strength as a concept). The assumption made above, that memory strength and maximum error are inversely related, is a relatively weak one. We would like to define this relationship more concretely to develop a useful model and interpret the distribution of errors in our task. To be clear, the maximum error (precision) P is some function of normally distributed memory strength S. The function $f(S)$ could take a number of forms, but should be monotonic: higher strength should always result in better precision, i.e. lower maximum error. Possibilities we tested empirically included relationships of linear and power form, but an exponential relationship, of the form $P = (\log(-S))$, provided the best fit. This exponential relationship means that improving precision (reducing error) when it is already good requires a much larger increase in memory strength than when precision is poorer. For example,

improving the maximum error from 2° to 1° requires a greater increase in memory strength than improving it from 20° to 10°. If memory strength is normally distributed, precision will follow a lognormal distribution (Figure 4.2(b)).

We can simplify this further, since the distribution of errors[2] implied by lognormally distributed precision, Figure 4.2(c), is closely approximated by a Cauchy distribution, Figure 4.2(d). A Cauchy distribution can be seen as more 'thresholded' than a Gaussian distribution in the sense that it comprises more very high and very low magnitude values, and fewer moderate ones (the ratio of two values drawn from a standard normal distribution, i.e. the t-distribution with one degree of freedom, is Cauchy distributed). The wrapped version of the distribution has the probability density function:

$$f(x;\gamma) = \sum_{n=-\infty}^{\infty} \frac{\gamma}{\pi(\gamma^2 + (x + 2\pi n)^2)} \tag{4.1}$$

We will analyse response errors directly using the Cauchy distribution throughout this chapter for several reasons. Firstly, the Cauchy distribution is simpler - it has a single scale parameter γ, while the lognormal-precision model uses two parameters (corresponding to the mean of the normal distribution, which controls the scale of the final error distribution, and the standard deviation, which controls its shape). Despite this, the more parsimonious Cauchy distribution provides a comparable (in fact slightly better) fit than the lognormal-precision model to the three sets of test data in this chapter. This may simply be a consequence of the fact that the standard deviation/shape parameter is matched across the three datasets (i.e. the final error distributions differ in scale but not shape), and is therefore superfluous in our data. Should this not hold in general, the shape of the distribution would change and a Cauchy distribution would not fit well. Secondly, a model defined in terms of the arguably ill-defined concept of 'memory strength' is less testable, less able to produce quantitative predictions and as a consequence less clearly related to mechanistic neural models than one defined in terms of measurable physical values, such as precision, error or accuracy. We believe greater progress can be made in future by moving beyond the concept of

[2]Note that in circular statistics, a single point on the circle corresponds to an infinite number of points on a distribution which is defined on the set of real numbers $\mathbf{R}$; i.e. n, $n \pm 2\pi$, $n \pm 4\pi \ldots$ Throughout this chapter, all error distributions are 'wrapped' around the circle in this way.

memory strength, in favour of more grounded terms. Nonetheless, we carried out the preceding analysis to demonstrate how the results we present in this chapter can be related to, or thought about in terms of, existing psychological models of memory.

4.1.3 Testing the presence of a threshold

Using the logic outlined above, we now have a clear prediction for the error distributions in the graded source memory task. Continuous and threshold models of recollection make conflicting predictions about the probability distribution of response errors expected (Figure 4.3). For continuous recollection, the relationship between underlying memory strength and the consequent error distribution on the task may take several forms (as we noted above). Briefly, however, the error distribution must be monotonically decreasing from low error to high error. This is because low strength trials can give rise to low errors (through chance) but high strength trials should not give rise to high errors. In contrast, if a threshold exists and recollection fails on a significant number of trials, these should be distributed evenly around the circle relative to the target (assuming that, in the absence of recollection, participants will be forced to guess). Thus for thresholded recollection, at low errors the distribution will decay rapidly (because of high-strength, recollected trials) but then stabilise to an asymptote which is greater than zero (because of zero-strength guessed trials).

We fit each participant's error distributions to a thresholded model of recollection, with two free parameters. A threshold parameter, λ, denotes the proportion of trials on which recollection succeeds. Errors on these recollected trials follow a Cauchy distribution with shape parameter γ: the spread of responses around the target, such that higher values of γ indicate a greater mean error. The remaining $1-\lambda$ non-recollected trials are guesses, randomly distributed around the circle relative to the target, resulting in a uniform distribution of error. When $\lambda = 1$ all responses are based on some, variable, recollection: no threshold is present, effectively rendering the model continuous. Alternatively, if $\lambda < 1$, recollection fails for a subset $(1-\lambda)$ of responses and a threshold exists.

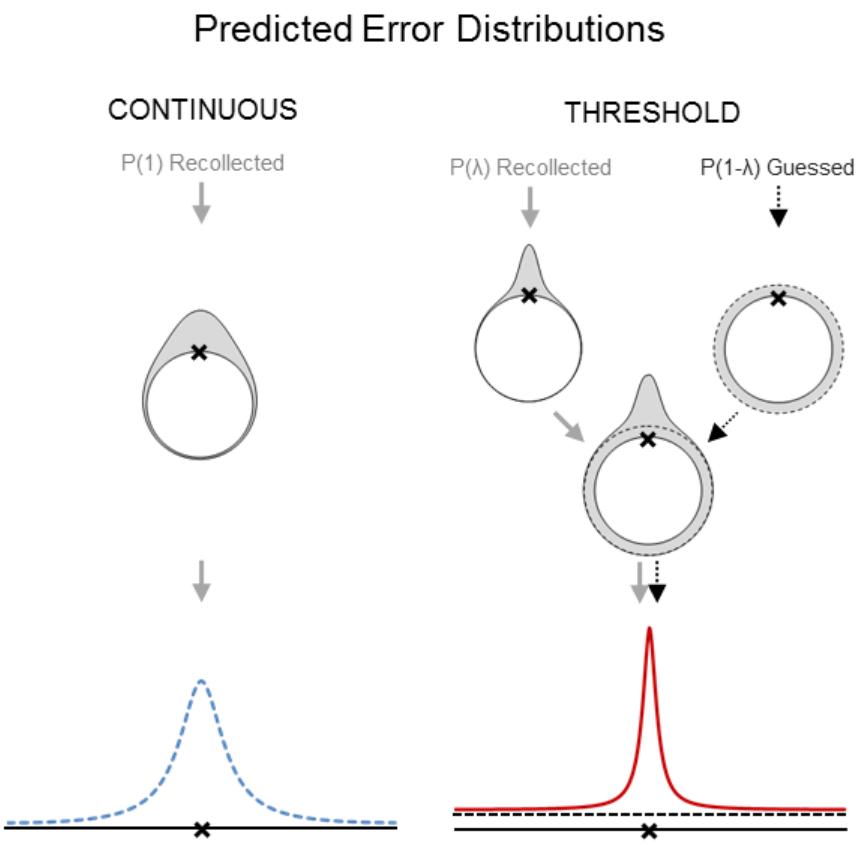


Figure 4.3: Predicted error distributions. Continuous recollection predicts a continuous error distribution, strictly decreasing away from the target (x). Thresholded recollection predicts a mixture of highly accurate successful recollection (λ) and sub-threshold guesses ($1-\lambda$) that shift the distribution asymptote above zero.

4.2 Experiment 1

In Experiment 1 we had two objectives. First, to test whether a threshold exists in the data; and second, to examine the relationship between accuracy and confidence. Twenty-seven participants (19 female; mean age 18.8, range 17-21) completed the experiment and all data sets were included in the final analysis.

4.2.1 Experiment 1 Methods

Experiment 1 followed the procedure outlined in Figure 4.1; details can be found in Chapter 3. Each participant completed 20 study-test blocks, 240 trials in total. In each block, participants studied 12 location/word pairs; attention to

each stimulus was verified by requiring participants to indicate the (now hidden) location with the mouse before continuing[3].

At test, participants were cued with each word from the preceding study phase (presented in a random order) and indicated their recalled location using the mouse, followed by a confidence judgment made by clicking along a near-continuous (600-pixel) scale (Figure 4.1(b)). No time limit was set for any response, and for each test phase response participants could change their selected location or confidence by clicking again elsewhere; a marker on screen indicated their current selection. To confirm their choice participants pressed a button on the serial response box, which recorded their currently marked selection and advanced the experiment to the next screen.

The distribution of errors (arc length between target and recalled locations, Figure 4.1(c)) was fit separately for each participant by maximum likelihood estimation to a mixture model of $(1-\lambda)$ sub-threshold guesses (uniform distribution) and λ recollected trials, modelled by a continuous wrapped Cauchy distribution (Figure 4.3). The significance ($\lambda < 1$) of this threshold was evaluated by a likelihood ratio test across the full dataset of 27 participants. Confidence ratings were used to form separate source ROCs[4] for each participant, which were fit to a mixture (i.e. graded and thresholded) model of source recollection.

4.2.2 Experiment 1 Results

Figure 4.3 shows the predicted mean distribution of response errors for continuous and thresholded models of memory; Figure 4.4 shows the observed error distribution, collapsed across all participants. These test data show that participants were well capable of performing the task: responses rates were highest close to the target location, indicating above-chance performance.

[3]Only the final responses on each trial were recorded in Experiment 1, these were therefore constrained to lie within 20 pixels of the target and we do not analyse them here. The initial response on each trial is more informative; these were recorded during Experiment 2 and are analysed there.

[4]Source ROCs were formed by randomly assigning each location to be a target or a lure (the choice makes no difference to the fit of the source ROC since it is constrained to be symmetrical around the secondary diagonal, $p(H) = -p(FA)$), and defining correct responses as those falling within 90 degrees of the true location, which makes performance analogous to that obtained using two equally-likely possible sources.

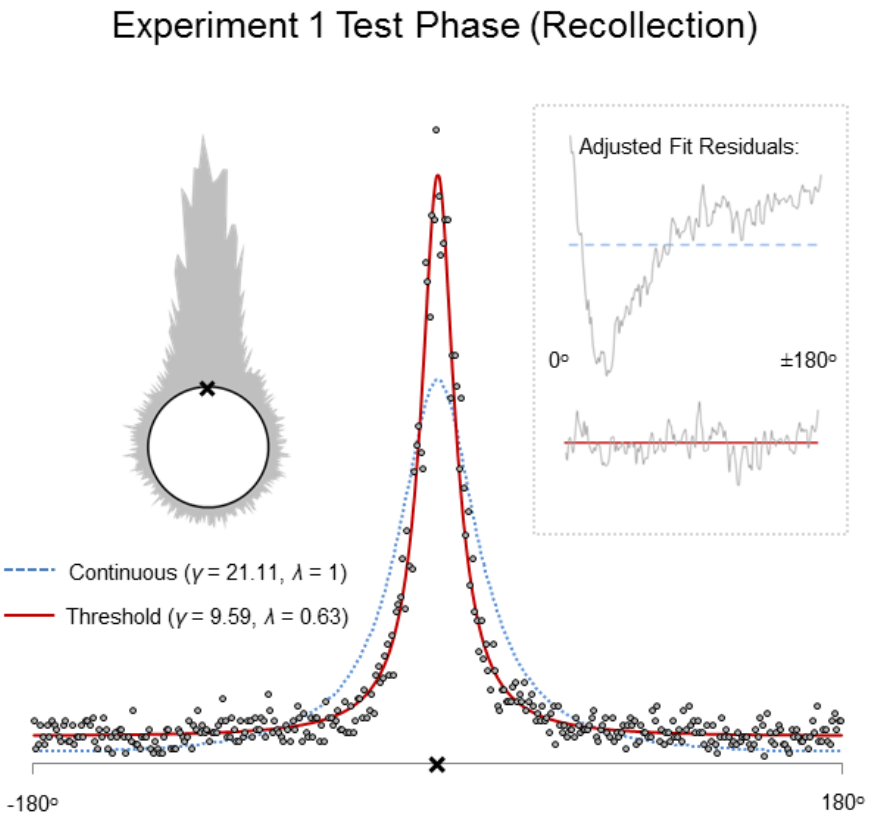


Figure 4.4: Observed error distribution, Experiment 1 (retrieval). Responses relative to the target (×) are shown around the circle, inset left, and unwrapped, centre. Thresholded (red, solid) and continuous (blue, broken) model fits are included with best-fitting parameters for the aggregate data. Adjusted fit residuals, inset right, show the observed data (grey) relative to each of the model predictions as a function of error magnitude. These highlight that the data systematically deviate from the best-fitting continuous model, but closely follow the prediction of the thresholded model.

We tested for the presence of a threshold by a likelihood ratio test: if a threshold exists, allowing λ to vary (the threshold model) should significantly improve the likelihood of the observed data compared to fixing it at $\lambda = 1$ (the continuous model). Allowing λ to vary dramatically and significantly improved the fit compared to the purely continuous model (mean $\lambda = .635, \chi^2(27) = 1153.30, p < .001$). To be clear, the thresholded model provides a greatly improved fit over the continuous model. We also tested goodness-of-fit using the G-statistic (Sokal and Rohlf, 1995), which confirmed that the thresholded model fit the aggregate data well ($\chi^2(177) = 169.91, p = .636$) but the continuous model did not ($\chi^2(178) = 1348.28, p < .001$). The adjusted fit residuals of each model (Student's t-statistic, Figure 4.4 inset) make the reason for this clear: the thresholded model fits consistently well across the entire range of errors, but the continuous model systematically underestimates the number of highly accurate and highly inaccurate responses. In essence, the continuous model fits poorly because the data is more thresholded than a continuous account predicts. Recollection is subject to a threshold: it fails completely on over a third of trials.

4.2.3 Confidence data

To facilitate comparison with previous studies we also acquired confidence data for each participant, allowing us to construct a symmetric source ROC curve for each participant. The group ROC and z-ROC are shown in Figure 4.5. It is notable that the ROC is relatively curvilinear (though on close inspection it is flattened in the centre, reflecting the presence of a threshold) and the z-ROC is close to being linear. Any subtle deviations from this pattern are especially difficult to detect when the data is binned into 6 points (Figure 4.6), which is a common level of definition in memory ROC studies (Yonelinas and Parks, 2007). In particular, analysing the significance of the quadratic component of zROCs at this or similar resolution (a test generally considered to reflect the presence or otherwise of a threshold, see Section 2.2.4 and, for example, Wixted, 2007a but also many other recognition memory studies) would provide no hint that the underlying memory data is not normally distributed and continuous, since the z-ROC deviates from linearity only very subtly. Even when a threshold is clearly visible in the accuracy data, it is much harder to detect it in confidence data.

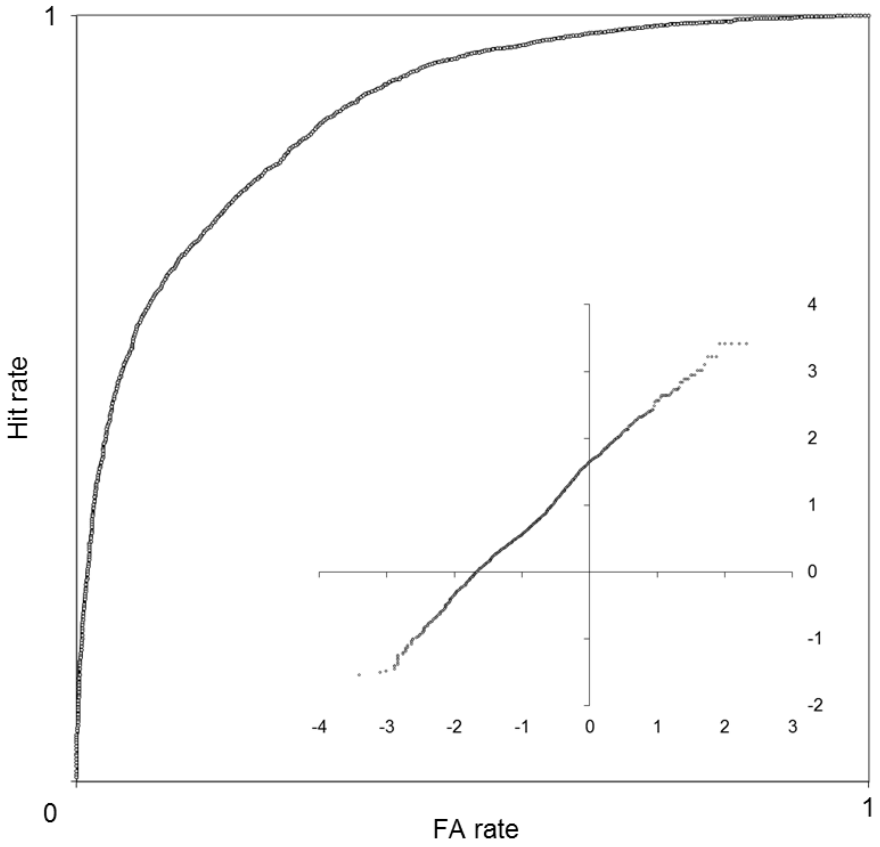


Figure 4.5: Fine-grained (600pt) ROC and z-ROC (inset), Experiment 1. The flattening in the centre of the ROC curve, and slight negative deflection in the centre of the z-ROC, reflect the presence of a threshold in the confidence data, but these effects are relatively subtle.

Does this mean that the threshold is not, in fact, present in the confidence data at all? Indeed, do subtle differences in memory really account for much of the variance in confidence judgments, or does the variance instead arise primarily from other non-mnemonic sources? In Figure 4.7 we examine the relationship between confidence and accuracy directly, by plotting mean error as a function of confidence rating (from 1-600, binned into 100 points and collapsed across participants). Confidence and accuracy appear approximately linearly related for successfully recollected trials, i.e. those on the right hand (higher-confidence) side of the graph. The relationship changes for lower-confidence responses, since these comprise significant numbers of guessed trials. In this part of the plot confidence is less closely related to accuracy or precision, since no diagnostic evidence is available to differentiate guess trials from each other. If all recollected trials were

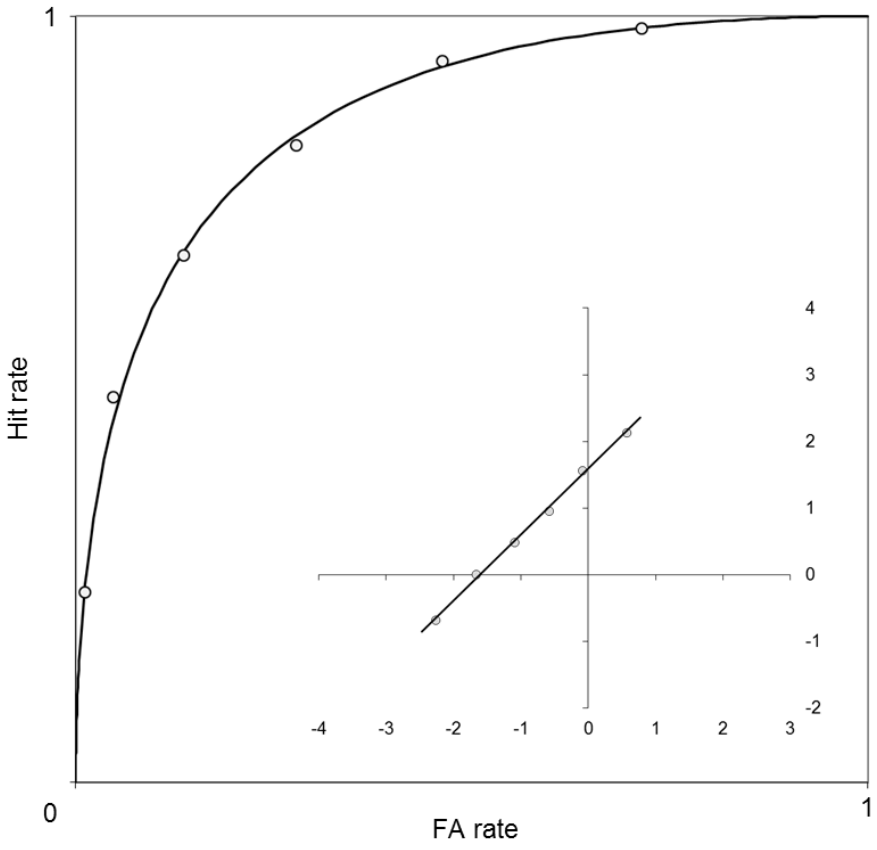


Figure 4.6: Coarse-grained (7pt) ROC and z-ROC (inset), Experiment 1. This resolution is comparable to the majority of existing ROC studies, and although a threshold is present the data is virtually indistinguishable from an entirely continuous account (solid line), highlighting the fact that confidence data are poorly suited to distinguishing between thresholded and continuous accounts of recollection.

assigned greater confidence than every guessed trial, as required by the DPSD model, this part of the plot would have zero gradient. To the extent that there is overlap in the distributions of confidence for guessed and recollected trials (e.g. because the data is collapsed across participants) the low-confidence, constant accuracy portion of the graph blends smoothly into the high-confidence, linear portion.

Figure 4.7: A linear relationship between confidence and accuracy for successful recollection. Each marker shows the mean distance from the target (with standard error) for a given confidence rating in Experiment 1, and the dashed line indicates chance performance (mean error of $90°$). The trendline is consistent with a linear relationship between error and confidence on recollected trials, such that participants' confidence ratings are on average around 1% higher for each degree closer they land to the target (when guessing, responses are drawn from a normal distribution of very low confidence). By this explanation, participants are able to distinguish trials of differing precision.

Thus, when recollection occurs it seems that participants' confidence ratings may in fact be highly sensitive to the precision of their memory, with confidence increasing by slightly more than 1 percentage point for every $1°$ closer to the target location they respond. The trendline in Figure 4.7 incorporates this relationship,

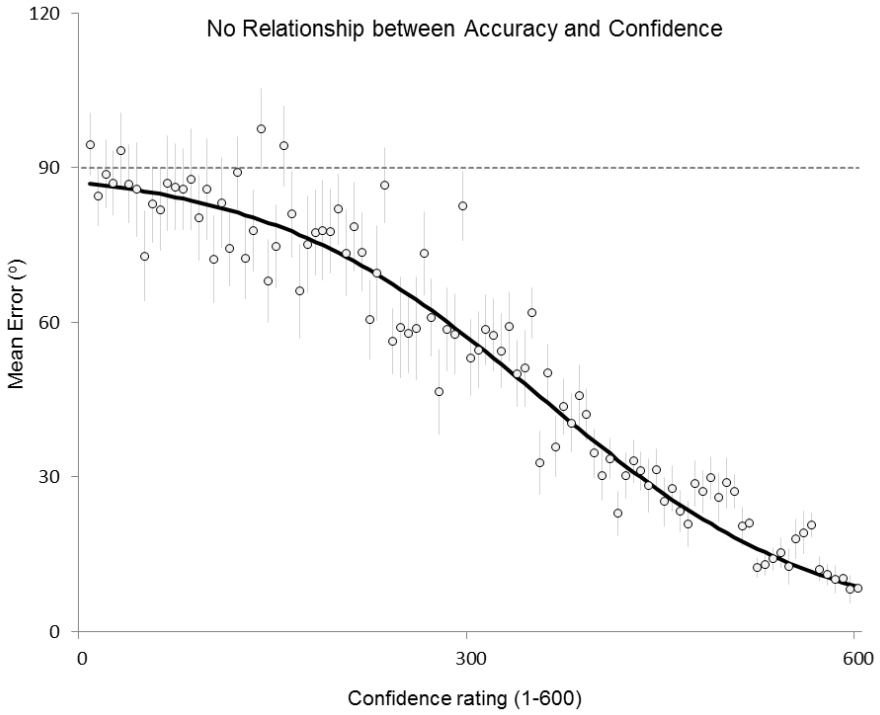


Figure 4.8: No relationship between confidence and accuracy. The trendline here includes no explicit relationship between error and confidence, and the pattern simply reflects a greater mean confidence rating to successfully recollected than guessed trials. By this explanation, participants are only able to distinguish between recollection success and failure - but not trials of differing precision.

together with some reasonable assumptions about the proportion of trials which are recollected (63.5%, i.e. the value of λ estimated from the accuracy data) and the distribution of confidence ratings for recollected and non-recollected trials (Gaussian).

Although this seems to be a persuasive account of the data, it is wise to be cautious before interpreting this pattern as evidence that confidence reflects graded memory strength (e.g. Mickes et al., 2009, Slotnick, 2010) since mathematically it can be explained in other ways. For example, the trendline in Figure 4.8 does a similarly good job of accounting for the relationship between confidence and accuracy, but in this case confidence is unrelated to memory precision. Instead the relationship is entirely explained by the fact that successful recollection yields higher average confidence than memory failure, but the two distributions overlap. Thus higher confidences have lower mean error since they contain fewer guesses, not because the recollected trials are more precise, and the data are consistent

with an all-or-none (in terms of confidence) model of recollection. The question of whether confidence judgments are sensitive to graded recollection strength depends on whether recollection strength is actually graded, or whether recollection is entirely determined by the success or failure of an all-or-none, thresholded process.

In Experiment 1 we have been able to demonstrate a clear threshold in the response data, highlighting in the process the advantage of using accuracy data to examine memory. Nonetheless, as we have emphasised previously (Section 2.3.5) this pattern does not in itself require that recollection is thresholded. A threshold in these data might equally be explained by intermittent attention or encoding during the study phase (though we took some steps to mitigate this, by requiring participants to reproduce the studied location). Moreover, the accuracy and confidence data further hinted that successful may also be detectably varied. The evidence for this latter conclusion is not conclusive either, since imprecision did not correlate with overall performance and may simply reflect motor or encoding errors, and the observed correlation between confidence and error might reflect changes in the proportion of recollected trials. We can establish whether the imprecision and threshold we observe in the data are related to memory by testing whether each property changes with study-test delay. If they do, this indicates that they reflect characteristics of retention or retrieval, and not factors which are invariant to study-test delay such as encoding failure or motor error. Thus, Experiment 2 was similar to Experiment 1 except that we tested participants after either a short or long delay, holding study conditions constant to remove potential confounds of encoding failure and non-mnemonic imprecision.

4.3 Experiment 2

The main aim of Experiment 2 was to determine how memory performance changed as a function of the time between study and test. Seventeen participants (11 female; mean age 19.1, range 17-24) completed the experiment and all data sets were included in the final analysis.

4.3.1 Experiment 2 Methods

Experiment 2 followed the procedure outlined in Figure 4.1; details can be found in Chapter 3. Each participant studied 36 blocks of nine word-location pairs each, 324 trials in total. The procedure was designed so that half of these blocks comprised a 'short delay' condition, in which participants were tested shortly after completing the study phase (just as for Experiment 1). For the remaining, intermixed, blocks, testing was delayed until two blocks later. To be clear, for this 'long delay' condition participants completed either 2 study blocks, 1 study and 1 test block, or 2 test blocks, before being tested on the long delay test block (and this was arranged so that participants could not predict, from the preceding blocks, whether any given study block would be in a short or long delay condition). To mitigate the chance that working memory may contribute to performance in the short delay condition, we removed all test trials that were separated from their corresponding study trial by fewer than 3 other trials plus the 10s gap before analysis.

4.3.2 Experiment 2 Results

To control encoding, participants were required to indicate each location during the study phase (Figure 4.1(a)). In Experiment 2, we recorded the initial response to this task on each trial. These study data (Figure 4.9) indicate that participants were very accurate at reproducing the target location within a few seconds of seeing it: 93% of trials were accurate within 10°. Interestingly, the attended study data more closely followed a Gaussian[5] than a Cauchy distribution (Table 4.1), which is arguably consistent with previous findings that working memory is associated with approximately normal distributions of error (Luck and Zhang, 2009). By contrast, at test, successfully recollected trials followed a wrapped Cauchy distribution. The qualitative difference between the distributions seen at study and test is consistent with the widely held view that working and long-term

[5]It is worth noting that while the Gaussian distribution provided a far superior fit to the study data than the Cauchy distribution did, it did not fit the data perfectly and was rejected by a G-test. Perhaps the working memory data comprised a more complex convolution of distributions, for example because some of the error arose from working memory and some from motor error. The sensitivity of the G-statistic and large number of observations in our study mean that we might have detected small deviations from normality that were not obvious in previous studies.

memory reflect the operation of different underlying memory systems (Atkinson and Shiffrin, 1968, Corkin, 2002), confirming that our novel source retrieval task relies on episodic memory.

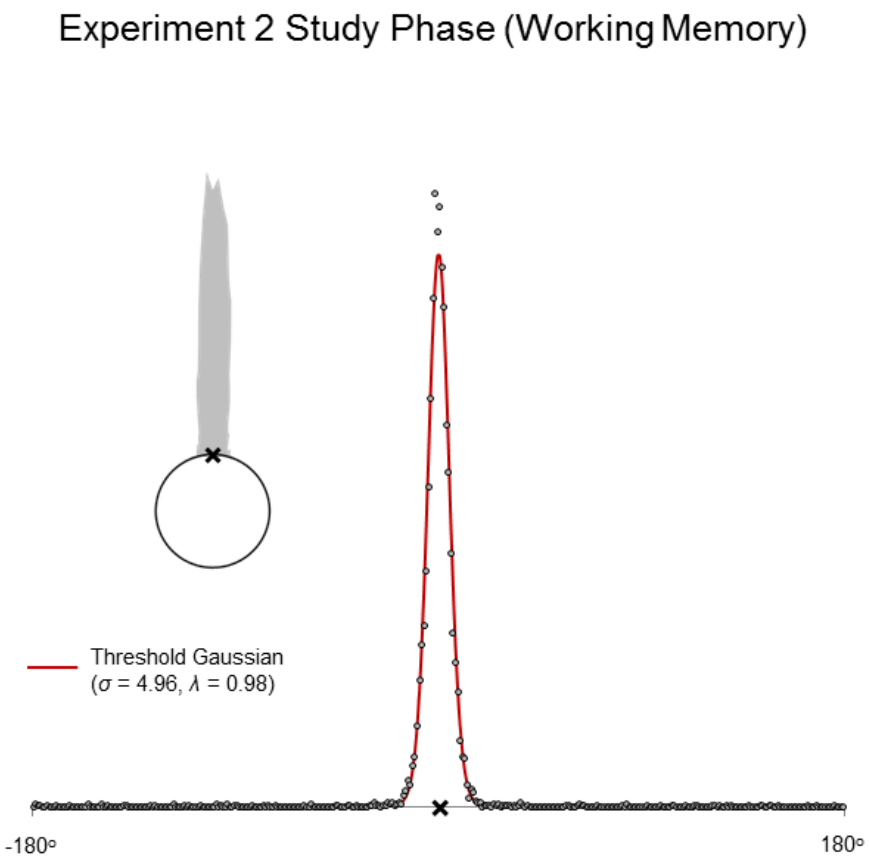


Figure 4.9: Observed error distribution, Experiment 2 (study data). At study, responses were highly accurate with 98% of trials attended, and 93% of trials landing within 10° of the target. Unlike responses at retrieval, the study data was more closely approximated by a Gaussian than a Cauchy distribution, possibly because motor error played a proportionally greater role.

The test data, shown separately for short (Figure 4.10) and long (Figure 4.11) delays, again demonstrate that participants performed well above chance, response rates were clearly highest at the target location. For both short and long test delay conditions, a threshold was required to explain the distribution of errors: allowing λ to vary significantly improved the fit compared to the purely continuous model (mean $\lambda_{short} = 0.698, \chi^2(17) = 376.68, p < .001$; mean $\lambda_{long} = 0.593, \chi^2(17) = 292.61, p < .001$). Just as in Experiment 1, goodness-of-fit

	Gaussian Model			Cauchy Model		
Data	LL	G	p	LL	G	p
Exp. 2 Study	-13712	323.3	<.001	-14140	1180.6	<.001
Exp. 1 Test	-30015	324.1	<.001	-29938	169.4	.645
Exp. 2 Test (short delay)	-11484	269.3	<.001	-11439	177.8	.468
Exp. 2 Test (long delay)	-13253	212.1	.037	-13230	164.2	.745

Table 4.1: Fit statistics for Gaussian and Cauchy models of Experiment 1 & 2 data. LL denotes log-likelihood of the data (lower magnitude indicates better fit), G denotes the G-statistic (lower values indicate better fit) and p $< .05$ indicates the model was rejected by a G-test. Errors from recollection at test were consistently well described by a Cauchy distribution, whereas responses at study more closely (but not perfectly) approximated a normal distribution. Both models included λ, i.e. they allowed for some proportion of trials to be guesses.

tests using the G-statistic confirmed that the thresholded model fit the aggregate data well ($\chi^2(354) = 328.46, p = .831$) but the continuous model did not ($\chi^2(356) = 1116.78, p < .001$), and the fit residuals (Figures 4.10 and 4.11 insets) showed that this reflected the same pattern that we observed in Experiment 1: the continuous model fit poorly because the data was thresholded.

Although the data is thresholded, proponents of continuous models have argued that recollection failure may occur simply because contextual source information has not been attended to or successfully encoded (DeCarlo, 2003, Mickes et al., 2010). Equally (though less frequently pointed out), variable accuracy for successfully recollected trials may not actually reflect variable degrees of recollection, but rather imprecise encoding at study. To rule out encoding factors as a confound we compared responses after short and long delays, which have identical encoding, but different retrieval delays. If the threshold is a product of encoding, the proportion of recollected trials, λ, should be matched for short and long delays. Similarly, if response variability simply reflects imprecise encoding or motor error, the scale parameter γ should also be independent of study-test delay.

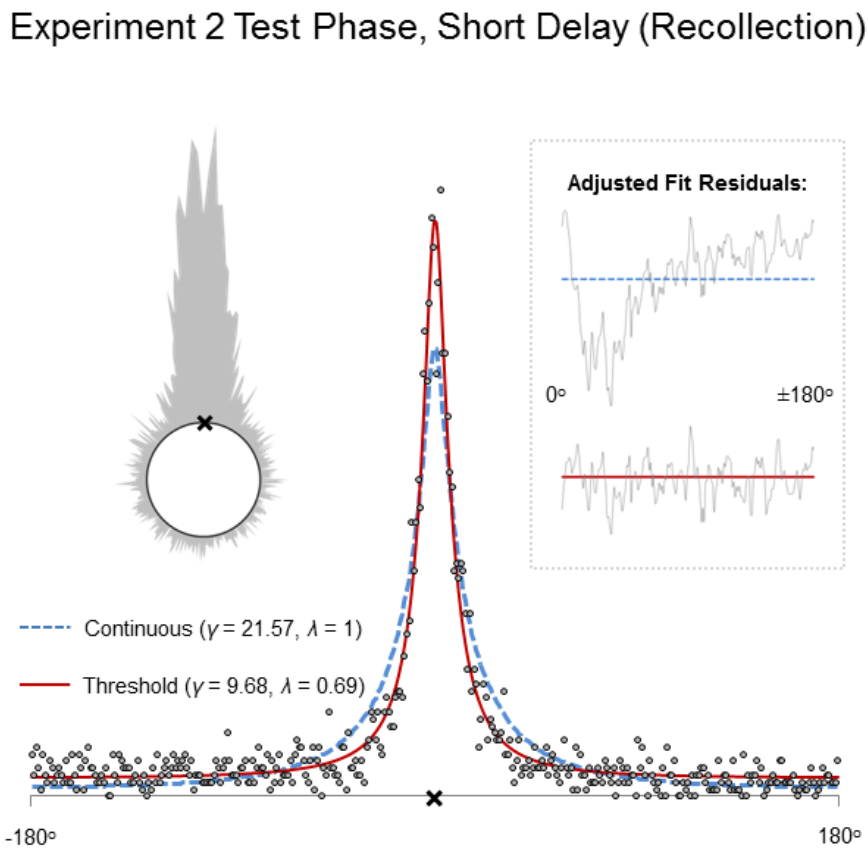


Figure 4.10: Observed error distribution, Experiment 2 (short delay). After a short study-test delay responses were thresholded: 69% were recollected (following a Cauchy distribution), and 31% were guesses. Adjusted fit residuals, inset right, once again demonstrate that the thresholded, but not the continuous, model account for the data pattern observed.

Paired t-tests indicated that in fact both λ and γ differed across conditions. After a longer delay recollection was less frequent (mean $\lambda_{short} = .69$, mean $\lambda_{long} = .59; t(16) = 4.02, p < .001$) and less accurate (mean $\gamma_{short} = 9.68$, mean $\gamma_{long} = 13.93; t(16) = 2.83, p = .012$). Furthermore, allowing γ to vary between short and long delay significantly improved the likelihood of the data ($\chi^2(17) = 72.33, p < .001$), indicating that successful recollection provided genuinely variable accuracy: it was graded. Allowing λ to vary between conditions also significantly improved the likelihood of the data ($\chi^2(17) = 38.71, p = .002$), confirming the presence of a retrieval threshold, which was independent of encoding.

We also ruled out one other subtle way in which a continuous model could account for the threshold. According to a source misattribution explanation (Slotnick,

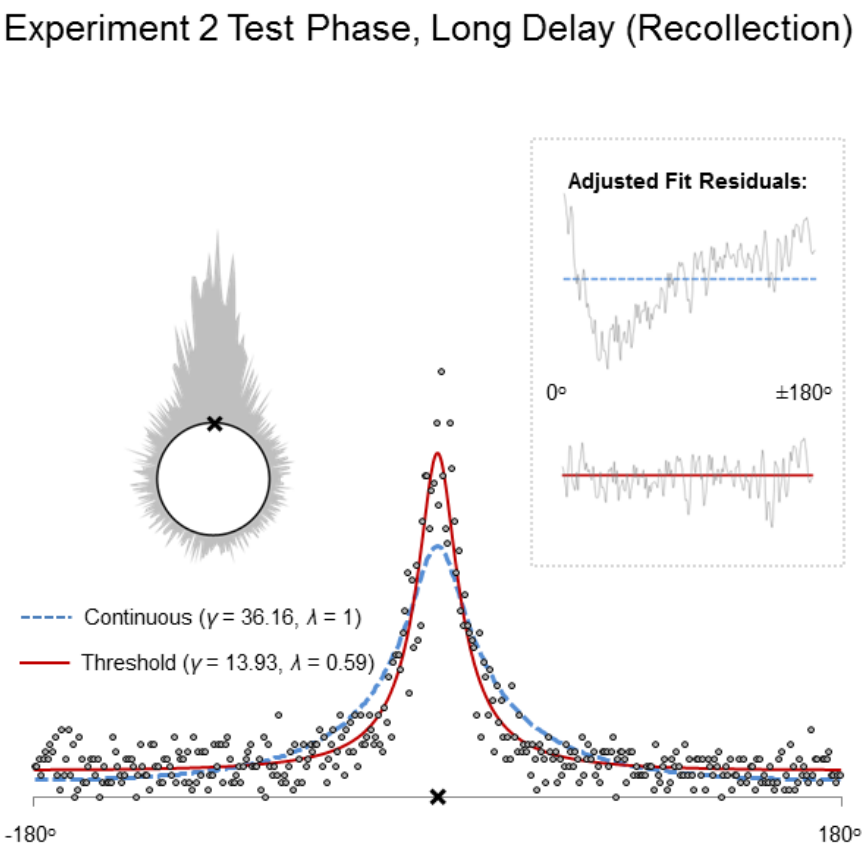


Figure 4.11: Observed error distribution, Experiment 2 (long delay). After a longer (2-phase, i.e. 18-trial) delay between the study and test phases, recollection was less likely to succeed (59%) and was also less accurate when it did, strongly supporting a graded and thresholded characterization of recollection.

2010), sub-threshold 'guess' trials, which provide the evidence for a threshold, may in fact have been misremembered (i.e., they reflected retrieval of another location from the study block). We tested whether this was the case by examining trials where the correct location had most likely not been recollected: responses more than 90° away from the target. Did they fall closer to any other locations from that study block than would be expected by chance? Errors were redefined as the distance between each response and the closest studied location from the same block, not the correct (target) location. These were then fit to the same model of guessing and recollection described above, with the exception that the guess rate was replaced with the distribution of errors expected by chance (enumerated exactly for each study block). Likelihood ratio tests confirmed that allowing the model parameters (γ, λ) to vary did not improve the fit compared to the guess

distribution alone ($\chi^2(2) = 1.04, p = .594$), i.e. the distribution was best fit by pure guessing. Incorrect trials were no closer to other locations than would be expected by chance and a continuous model cannot, therefore, account for the threshold.

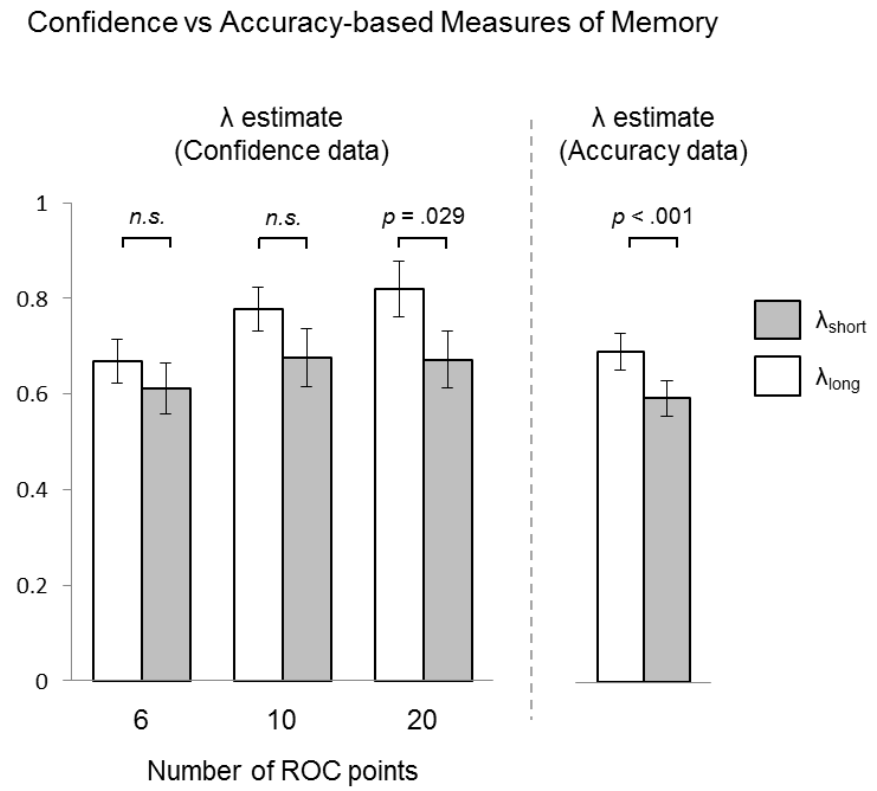


Figure 4.12: Estimates of the critical threshold parameter λ, Experiment 2. Confidence provides reasonably accurate parameter estimates when the correct (variable and thresholded) model is used, but is underpowered for characterising recollection itself. A reduction in λ with study-test delay is evidence of a recollection threshold, but this difference is noisy in the confidence data, and statistically invisible in ROC curves of commonly-used resolutions (6-10 points). In contrast, the source accuracy task eliminates the need to model the relationship between memory strength and confidence for each participant, resulting in more stable parameter estimates. Using accuracy, but not confidence, recollection can be clearly characterized as thresholded.

Finally, to demonstrate the practical effects of using ROC curves to examine memory, we again constructed symmetric source ROC curves using the confidence data, this time with the aim of testing whether the critical result found above - that the threshold varied independently of encoding - was still visible.

Importantly, we found that this difference was only visible in ROCs with a far higher resolution than are generally used (e.g. 20 points: $\lambda_{short} = .82, \lambda_{long} = .67; t(16) = 2.40, p = .029$). When confidence ratings were binned to form 6 or 10 point source ROCs (the resolution employed in the overwhelming majority of previous ROC studies, e.g. DeCarlo, 2003, Eichenbaum et al., 2010, Harlow et al., 2010, Hilford et al., 2002, Howard et al., 2006, Mickes et al., 2010, Onyper et al., 2010, Peters et al., 2009, Sherman et al., 2003, Slotnick, 2010, Slotnick and Dodson, 2005, Yonelinas, 1994; 1999), there was insufficient power to detect a change in the rate of recollection (Figure 4.12; see also Figure 4.6 for an illustration of ROC data at this resolution). The relatively low sensitivity of confidence data to detect a threshold amidst the type of noise we discussed in Section 2.3.5 may help to explain why previous attempts to characterise recollection using ROCs have provided conflicting evidence. In short, direct comparison suggests that memory is much more accurately characterised by objective measures of memory strength such as error than by subjective measures such as confidence.

4.3.3 Which determines memory performance: a threshold or graded recollection?

The threshold model we have fit here is not all-or-none: errors can be introduced both through memory failure and imprecision. How important are each of these effects? We tested this by measuring the mean error in degrees for each participant, and calculating how much of this was accounted for by memory failure, using each participant's estimate of λ and assuming a mean error of 90° on guessed trials. In Experiment 1, participants' overall mean error was 46.5°, and 32.5° of this on average was introduced as a result of sub-threshold guessing, as opposed to 14.0° through imprecision. Thus, even though the estimate for memory imprecision can be considered an upper bound, since it includes motor error, the proportion of trials for which recollection occurs is still more critical to overall performance ($t(26) = 8.28, p < .001$). Viewed another way, λ explains virtually all of the variance in overall performance (mean error): the two are strongly correlated ($r(25) = -.96, p < .001$). In Experiment 2 we observe a similar story, although in this case we can also estimate the effect of motor (and other non-mnemonic) response errors. We operationalized these as the mean er-

ror observed during the study task, i.e. when memory failure and imprecision should not be significant, which was 5.4°. For trials tested after a short delay, mean error was 42.1°, with 27.9° of this resulting from sub-threshold guessing and only 8.8° through memory-related loss of precision ($t(16) = 5.09, p < .001$). Similarly, for trials tested after a longer delay the mean error was 53.0°, with 36.6° of this resulting from sub-threshold guessing and only 11.0° through memory-related loss of precision ($t(16) = 6.32, p < .001$). The rate of recollection, λ, was strongly correlated with overall performance in both cases ($r(15) = -.97, p < .001$ and $r(15) = -.84, p < .001$ respectively) whereas memory precision γ was not ($r(15) = -.00, p < .997$ and $r(15) = -.33, p = .196$). Performance on even this task, which is highly sensitive to the precision of recollection, was almost entirely determined by the rate - and not strength - of recollection.

4.3.4 Why use the Cauchy distribution?

Here we have used a Cauchy model to fit the data and draw conclusions. How dependent are the results on this choice, and how sure can we be that it represents the appropriate model for our data? Although it is impossible to exhaustively test every possible alternative distribution, we can make the claim that the Cauchy distribution fit well: G-tests on each of the three sets of test data presented in this chapter failed to reject the model, Table 4.1, despite relatively high power (over 10,000 responses total). Furthermore, we can compare the fit to two representative alternatives, a Gaussian model which is less peaked and has lighter tails and a Pareto distribution which is more peaked and has heavier tails. These characteristics are relevant: as the weight of the distribution tail increases, it can capture more of the inaccurate responses and therefore reduce the estimated guess rate. In essence, the more peaked and heavy-tailed a distribution is, the more 'thresholdedness' in the data is assigned to the continuous part of the model. To test how well each model accounted for the data, Bayesian Information Criteria (BIC) were calculated for each and compared. BIC is a function of likelihood that corrects for the number of free parameters in a model; a lower value of BIC indicates a superior fit. Where L is the likelihood of the data under a given model and its estimated parameters, k denotes the number of free parameters in the model, and n is the number of observations then:

$$BIC = -2\log(L) + k\log(n)$$

As can be seen from the aggregate fit statistics in Table 4.2, however, the Cauchy model provides a far superior fit to both more and less thresholded alternatives. This bolsters the conclusion that the Cauchy model is a reasonable fit to the data.

Dataset	Gaussian BIC	Cauchy BIC	Pareto BIC
Exp. 1	156	0	932
Exp. 2 (short delay)	1892	0	418
Exp. 2 (long delay)	46	0	396

Table 4.2: Relative BIC for Gaussian, Cauchy and Pareto models of test data in this chapter. Responses were aggregated across participants for each dataset and fit to the three models. Bayesian Information Criteria were calculated for each model, and a constant (equal to the smallest BIC value for that dataset) was subtracted from each for clarity of comparison. These demonstrate that the Cauchy distribution was a consistently superior fit to the alternative models. All three models included λ, i.e. they allowed for some proportion of trials to be guesses.

Perhaps most importantly, however, using a different distribution will not change the key finding of the study: that a threshold is required to explain the pattern of errors made on our task. This is because the distribution of memory strengths or precisions across trials determines the slope of the error distribution observed. This is actually quite a complex relationship, linked to the fact that the error distribution must monotonically increase towards the target. The monotonic increase arises because both precise and guessed trials - the latter through chance - can land close to the target, but precise trials cannot land far from the target.

The Cauchy distribution of error that we observe implies a skewed strength distribution, such that most trials are relatively precise, but there is a 'long tail' of less precise trials. The greatest slope in the error distribution (around 2°-10° error) implies a corresponding peak in the number of trials with this level of precision. Where the slope of the error distribution is zero, the corresponding value of the strength distribution must be zero. Thus, no trials have precision 0 (infinite

strength) since the slope exactly at the target is zero. Similarly, the error slope is effectively zero from around 90° to 180°, meaning that there are no trials with a precision in that range. Critically, however, there are large numbers of trials with precision worse than this: 30% to 40% of trials must have a precision worse than 180°, i.e. they carry no information about where on the circle the target is located. These show up in the error distribution as an additive constant, causing a step function (infinite slope) at exactly 180°.

What this means is that the distribution of strengths cannot be a single continuous function. It has a peak at very high strength (a large proportion of the locations are recalled to within 10°, close to the baseline motor and working memory error we measure from the study data). Conversely, there are virtually no trials for which participants have a relatively imprecise knowledge of the location (e.g. the correct half of the circle). Finally, there is a further group of trials which have effectively zero strength. In other words, the strength distribution is bimodal, with peaks at both very high strength and zero strength. Any function which was designed to fit this bimodal pattern (to replace the Cauchy plus guessing model we use, which fits the data extremely well) would therefore effectively be a threshold model presented in different language. This thresholded pattern, where responses are either very accurate or guessed trials, with none of moderate strength in between, is inherent to the data and does not appear as a consequence of the function used at analysis.

4.4 Discussion

The ability to recollect individual events is fundamental to episodic memory, yet the nature of recollection has long been disputed. Here, by using a novel test of source accuracy, we are able to clearly characterize recollection: it exhibits a threshold and sometimes fails, even when the information being sought was successfully encoded, but it is graded when it succeeds. This characterization can explain previous findings in which attention or encoding effects could not be ruled out in favour of a retrieval threshold (DeCarlo, 2003, Hautus et al., 2008, Mickes et al., 2010). More importantly, the present study demonstrates that the threshold is, in fact, a real property of episodic recollection.

Accurately describing the way in which recollection operates is of considerable practical importance because the interpretation of results from lesion, imaging and other memory studies all ultimately rely on how memory processes are modelled. The fact that recollection has a threshold allows it to be objectively measured and separated from other memory processes, supporting an extensive body of research which continues to illuminate the function (Elfman et al., 2008, Greve et al., 2007), neurobiology (Eichenbaum et al., 2010, Peters et al., 2009) and decline (Howard et al., 2006) of episodic memory. Despite ruling out a continuous account of recollection, it is important to highlight that our findings are in agreement with others showing that when it does occur, recollection yields information of varying quality (DeCarlo, 2003, Hilford et al., 2002, Mickes et al., 2010; 2009, Onyper et al., 2010, Slotnick, 2010, Slotnick and Dodson, 2005). One alternative possibility, of course, is that recollection itself is not graded, but that instead successful recollection triggers the engagement of a graded process. For example, recollection might provide an approximate location (whose precision does not appreciably vary across trials), which participants then narrow down by repeatedly using a global-matching signal, such as familiarity.

Whether or not it does so directly or by triggering additional processes, however, it is clear from these data that recollection can ultimately give rise to variable information. Drawing process estimates from models which treat recollection as being thresholded but not variable, as we and others have previously done (Diana et al., 2008, Eichenbaum et al., 2010, Harlow et al., 2010, Haskins et al., 2008, Howard et al., 2006, Peters et al., 2009), will underestimate the contribution of recollection. To be accurate, future studies must address the variable nature of recollection explicitly by characterising it as a graded, thresholded process.

It is important to make the point that while we find a very clear recollection threshold in these data, we have done so using a task which has been carefully designed to address this specific question. Thus, it is possible that in tasks which depart significantly from this design (and could be argued to better reflect normal everyday memory) a recollection threshold may not be so easy to detect. For example, complex stimuli may allow partial recollection, or confidence may depend on factors other than the precision of recollection. We believe that these are strengths of our task, since our primary aim is to determine the properties of recollection (and therefore to constrain biological, cognitive and computational models

of memory), and not to determine the properties of confidence ratings (which are a relatively complex, impure and indirect reflection of memory). Nonetheless, since confidence ratings are frequently used as a measure of memory strength in practice, an important follow-up question is whether these results do really license the use of a graded and thresholded model of memory, such as DPMSD, to analyse confidence data. We believe they do, for several reasons. Firstly, our analysis of confidence ratings in this chapter strongly suggests that it is indeed legitimate to expect that memory properties should be reflected in confidence. Not only do the source ROC data follow a graded and thresholded mixture model, but the correlation between confidence and accuracy shown in Figures 4.7 and 4.8 seem likely to reflect a genuine sensitivity to recollection strength given that recollection is graded and not all-or-none. Secondly, a thresholded pattern has been found to fit confidence ratings well for source and associative memory tasks (e.g. DeCarlo, 2003, Hautus et al., 2008, Mickes et al., 2010; note that these studies disagree with the dual-process interpretation of the pattern which the DPMSD model carries, but crucially they find strong evidence for the pattern predicted by the DPMSD model). Finally, at the end of the next chapter we shall address this question directly by fitting different single and dual-process models to confidence data; to anticipate, the best-fitting models were those which incorporated both a recollection threshold, and variability in recollection strength.

The present findings raise the question of when the threshold arises, given that it is not determined at encoding. Do stored memory traces 'burn out' in a probabilistic way as new memories are stored? Or, are the computational mechanisms underlying recollection based retrieval intrinsically thresholded? One way in which these questions can be addressed is by bridging the predictions of neural network models with empirical data. The characterization of recollection proposed here is in striking agreement with the bi-modal distribution of recollection strength observed in neural network models (Elfman et al., 2008, Greve et al., 2010). For example, biologically inspired modelling suggests that the thresholded nature of recollection might reflect specific properties of the hippocampal network (Norman and O'Reilly, 2003). Intriguingly, however, computational modelling also demonstrates that the presence of a threshold may depend on the way in which a retrieval network is interrogated, such that the same memory representation can give rise to either a continuous or thresholded signal (Greve et al., 2010). By this

view, the mechanism(s) underlying recollection is itself inherently thresholded, not the representations upon which it operates. More broadly, retrieval processing may operate within the context of support processes that can themselves influence whether recollection succeeds (Rugg and Allan, 2000). For example, electrophysiological evidence shows that the adoption of an appropriate retrieval orientation is predictive of recollection success (Herron and Rugg, 2003). One important implication of this view is that when recollection fails, the memory being sought may nonetheless be retrievable using a different cue or by an alternative process such as familiarity.

In practice, of course, memories are likely to become inaccessible in multiple ways, for example by the corruption of traces beyond some minimum integrity or the gradual replacement of individual traces by new experiences. A key future direction for memory research should therefore be to establish, at the level of individual episodes, why recollection fails. Can memory deficits associated with aging and disease be isolated to particular supporting processes? Future studies must also investigate the constraints that recollection failure has for future retrieval: How does recollection failure influence the original memory trace? Is a delay required before subsequent attempts at recollecting can succeed? Is later recollection enhanced by the addition of noise or new signals to the memory network, or must an episode be approached using a different cue? Since recollection deficits are strongly associated with aging and dementia (Howard et al., 2006, Jennings and Jacoby, 1997, Kopelman, 1989, Naveh-Benjamin, 2000) the answers to these questions are likely to have an important impact on approaches to treating memory-related disorders - as well as understanding the vexing problem of locating one's keys.

In the following chapters we shall investigate an important consequence of the unreliability of recollection. The crucial ability to retrieve episodic associations - such as the name of a new acquaintance, or where we met them - has long thought to be dependent upon recollection. Perhaps, however, familiarity can also be used to retrieve this information, allowing us to mitigate the effects of increasing recollection failure as we age.

Chapter 5

Testing the Domain Dichotomy theory

Sections 5.1–5.4 of this chapter were published, with some minor differences, in the Journal of Experimental Psychology: Learning, Memory & Cognition (Harlow et al., 2010). These sections are followed by a subsequently revised analysis of the ROC data (Section 5.5), which is motivated by the conclusions of the previous chapter; namely, that recollection is both graded and thresholded.

5.1 Introduction

Dual-process theory posits the existence of familiarity and recollection, two functionally and neurally separable processes underlying episodic memory retrieval (for a review see Yonelinas, 2002a). An item is familiar if it simply engenders a sense of having been encountered before, whereas recollection provides additional contextual details about a previous episode. The two processes have been repeatedly dissociated, using a wide range of encoding conditions (Jacoby, 1998), retrieval tasks (Lecompte, 1995), and stimuli (Ratcliff et al., 1994), providing strong support for the dual-process distinction. Despite this, many substantive issues remain unresolved, including the relationship between the processes (Jacoby, 1991, Joordens and Merikle, 1993) and how they interact with other memory systems (Greve et al., 2007, Yovel and Paller, 2004). Here we focus on a related question: under what circumstances can familiarity contribute to successful recognition?

Familiarity is generally agreed to play an important role in standard item recognition memory tests, which assess memory for individual stimuli. Even when recollection is clearly impaired, for example in amnesic patients (Holdstock et al., 2002, Mayes et al., 2002), familiarity provides a strong basis for accurate performance. In contrast, in tests requiring memory for relationships between items, familiarity has traditionally been thought to play a less prominent role (Hockley and Consoli, 1999). Indeed, associative recognition tasks have been used to isolate recollection (e.g., Donaldson and Rugg, 1998), consistent with the belief that memory for such relationships should be supported exclusively by recollection (Yonelinas, 1997).

5.1.1 Does familiarity support associative recognition?

More recently, episodic memory theorists have begun to consider circumstances under which familiarity might contribute to associative recognition. In particular, a growing body of evidence suggests that when distinct stimuli are unitized (encoded and retrieved as a single unit) familiarity does contribute to associative recognition (Haskins et al., 2008, Quamme et al., 2007, Rhodes and Donaldson, 2007). For example, behavioural and imaging data suggest that pairs of linguistically associated words, like 'traffic-jam', evoke more familiarity at retrieval than semantically related word pairs, like 'cereal-bread' (Rhodes and Donaldson, 2008). Accordingly, some models of episodic memory propose that familiarity can support associative recognition, but only when to-be-remembered pairs are unitized (Diana et al., 2007, Eichenbaum et al., 2007).

Whilst unitization has received substantial empirical support, the 'domain dichotomy' theory (Mayes et al., 2007) provides an alternative account of why familiarity might sometimes contribute to associative recognition. According to this view, familiarity can support successful associative recognition even when stimuli are not unitized; instead, the contribution of familiarity is driven primarily by overlapping component representations in the medial temporal lobes. It is important to note that while item familiarity can support associative recognition indirectly (e.g., by providing a cue for recollection), both the unitization and domain dichotomy accounts propose and refer to a separate global familiarity for the associated pair. Here we provide a brief overview of domain dichotomy and

its empirical predictions, before presenting two experiments that directly test the domain dichotomy view.

5.1.2 The domain dichotomy theory

Domain dichotomy is based on a neuroanatomical account of medial temporal lobe function. At the heart of the theory is the separation of 'within-domain' and 'between-domain' associations. Within-domain associations (e.g., between two images or two words) occur between pairs of items that share some characteristics (e.g., modality; semantic category; component features) and are therefore likely to be represented by activity in overlapping populations of neurons in the perirhinal cortex. Between-domain associations (e.g., between an image and a word) conversely share fewer characteristics and so their representations are expected to be more distal and weakly connected.

This neuroanatomical account is itself derived in part from neural network models, which provide specific predictions about the role of familiarity. Computational models of familiarity typically invoke Hebbian type learning rules, causing similar inputs to be stored as similar patterns of activation and strengthening the overlap of these representations through repeated activation (Norman and O'Reilly, 2003) (but see Greve et al., 2010). This view implies that similar items should interact strongly, leading to better support from familiarity (Mayes et al., 2007). Consistent with this, some studies have shown patients with hippocampal lesions to be more strongly impaired at recognising between-domain than within-domain pairs (Mayes et al., 2004, Vargha-Khadem et al., 1997). Here we use healthy participants to test a prediction that domain dichotomy derives from lesion data, namely that within-domain pairs should be better supported by familiarity than between-domain pairs.

5.1.3 Testing domain dichotomy

We assess the predictions of domain dichotomy by examining associative recognition memory using two different measures of familiarity - safeguarding against the particular assumptions associated with each. First, we use confidence judgments made at test to form receiver-operator characteristic (ROC) curves; this

allows estimates of familiarity and recollection to be derived using mathematical memory models (see for example Yonelinas and Parks, 2007). Second, we use phenomenological data, asking participants directly about their memory experience. In the original remember-know procedure (Tulving, 1985) participants were required to identify if they recollected some aspect of the original experience (remember), or if they simply found the test stimulus familiar (know). Given recent criticism of this method, in particular by proponents of domain dichotomy (Mayes et al., 2007, Montaldi et al., 2006), here we use their modified procedure - making the terms familiarity and recollection explicit, training participants to distinguish recollection from high-confidence familiarity, and examining familiarity and recollection in separate tasks.

To examine memory we use a standard associative recognition task, presenting pairs of stimuli at study, and requiring participants to distinguish intact from rearranged pairs at test. If familiarity does contribute to successful associative recognition, both ROC analysis and the modified remember-know procedure should find evidence of it. Importantly, if the domain dichotomy view is correct, both methods should find greater estimates of familiarity for within-domain than between-domain pairs. As we explain below, both experiments found evidence of familiarity, but in stark contrast to the predictions of domain dichotomy theory it contributed more when pairs were between-domain.

5.2 Experiment 1

We employed names and abstract images as stimuli; because they differ both conceptually and perceptually they should occupy different 'domains'. In particular, each class of stimulus was chosen so that individual exemplars shared many features (e.g., size and shape), whilst still being individually distinguishable. Given these constraints, on average a name-image pair should be more between-domain than either a name-name pair or image-image pair. We also considered that one class of stimulus might be inherently more recognisable than another. To isolate relationship-driven memory differences, name-name and image-image pairs were collapsed to form a general within-domain condition, hence the relationship between items differed across conditions (within-domain; between-domain) but the items did not. It has also been suggested that two representations must be

directly encoded for overlap to occur (Mayes et al., 2007). We encouraged direct encoding in two ways. First, the items comprising each pair were presented simultaneously at study. Second, participants were instructed to judge how well the two items went together, without linking them via additional self-generated cues.

5.2.1 Experiment 1 Methods

In Experiment 1 we examined memory using 9-point ROC curves, constructed separately for each participant. We initially fit the data to a DPSD model of recognition; additional fits to alternative models are provided at the end of the chapter as a supplementary analysis.

Thirty right-handed participants completed Experiment 1; one data set was excluded due to non-compliance leaving a final cohort of 29 participants (11 female; mean age 22.8, range 18-31). Each stimulus comprised a pair of items presented above and below central fixation, as illustrated in Figure 3.2. We employed three stimulus conditions. Within-domain conditions comprised pairs of either Christian names (*WD-Names*) or abstract images (*WD-Images*); a between-domain (*BD*) condition comprised equal proportions of image-name and name-image pairs. In total 324 names and 324 images were used, see Section 3.1.2 for details.

The experiment was divided into 12 blocks, 4 for each stimulus condition, ordered randomly. Each block was further divided into a 27-trial study phase and an 18-trial test phase. At test, 9 pairs of items were intact (appeared together in the preceding study phase) and 9 were rearranged (appeared in separate study trials). For example, given three pairs A-B, C-D and E-F at study, an intact test pair would be A-B and a rearranged test pair C-F (discarding items D and E). Thus, every item shown at test had been encountered exactly once at study and successful performance required participants to remember the relationships between items.

Figure 5.1 shows the procedure for Experiment 1. Each study trial began with a blank screen for 500ms, followed by a central fixation cross for 1000ms, and a second blank screen for 100ms. The to-be-remembered pair was then presented for 3000ms. Following a 500ms blank screen participants were required to indicate

on a scale from 1-5 how well the two items went together; this response initiated the beginning of the next trial.

Test trials were identical to study trials except that each pair was presented for 1000ms, and the response screen asked participants to judge whether the items were intact or rearranged. Following the intact/rearranged response participants indicated how confident they were that they were correct, again using a scale of 1-5. This confidence response initiated the beginning of the next trial.

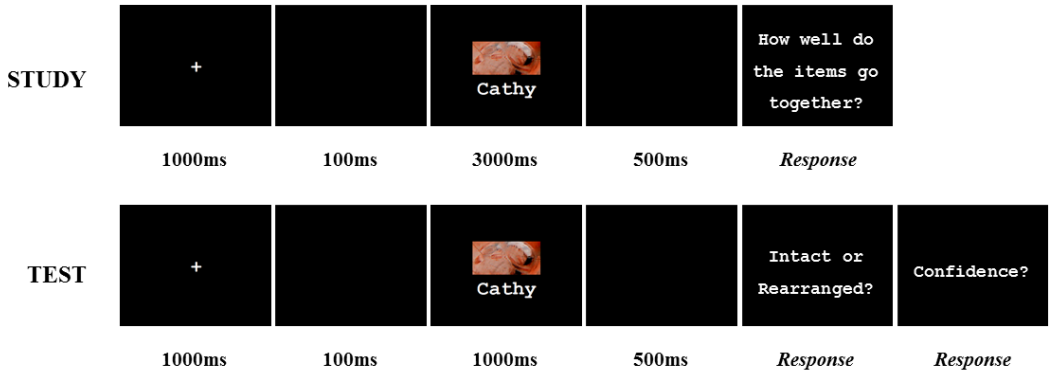


Figure 5.1: Experiment 1 procedure. At study, participants were presented with a pair of items and given a direct encoding task. At test, participants were presented with two items from the study phase and asked to judge whether they were originally presented together (intact) or in different pairs (rearranged), then indicate how confident they were of their answer on a five-point scale. Participants performed the task separately for pairs of names, pairs of images and mixed pairs.

At both study and test the mapping of left and right buttons to (*intact/rearranged*) and (1-5) responses was fully counterbalanced across blocks of 4 participants; the stimulus condition (*WD/WD/BD*) and test condition (*intact/rearranged/not shown*) of each item was fully counterbalanced across blocks of 9. On average the procedure took 1.5 hrs to complete, including a practice block and debriefing.

5.2.2 Experiment 1 Results

Mean ROC curves for each condition are presented in Figure 5.2; each exhibits clear curvilinearity, implying that either familiarity contributed to the associative recognition task or that recollection was graded. Below we explicitly assess

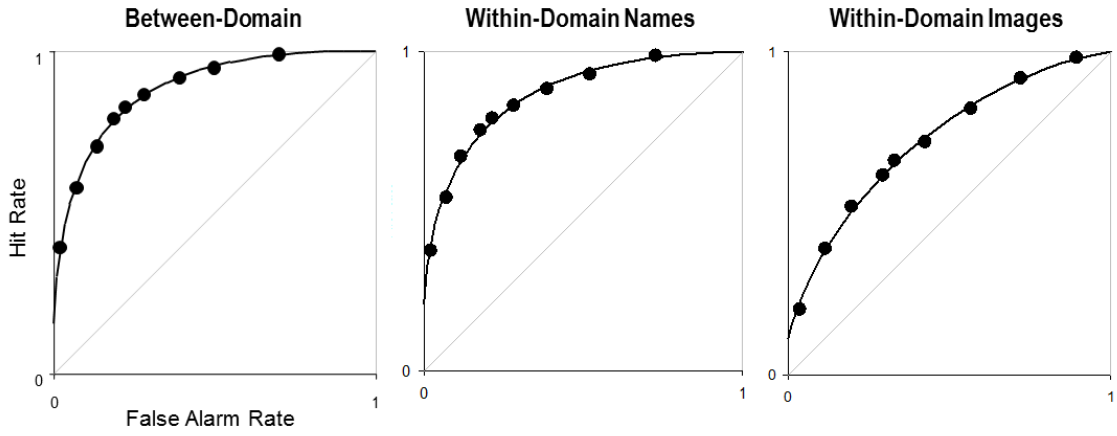


Figure 5.2: Group ROC curves for each condition from Experiment 1. Datapoints show mean hits and false alarms for each decision criterion; curves are from the best-fitting DPSD model in each case. Note that reported parameter estimates were obtained by fitting the DPSD model to individual participant data; these group ROCs provide a visual comparison between conditions. All three show clear curvilinearity.

whether the contribution of familiarity varies across conditions, as predicted by the domain dichotomy theory.

Overall task performance, mean familiarity estimates and recollection rates (collapsed across test condition) for each pair type are summarised in Figure 5.3. WD-Image pairs showed reduced discrimination (0.86) relative to BD (1.91; $t(28) = 6.62, p = .001$) or WD-Name (1.86; $t(28) = 6.19, p = .001$) pairs. No difference in discrimination was found between the BD and WD-name conditions ($p = .766$). Similarly, WD-Image pairs were less familiar (0.44) than either the BD (1.15; $t(28) = 4.00, p = .001$) or WD-Name (1.08; $t(28) = 3.69, p = .001$) pairs; WD-Name and BD again did not reliably differ ($p = .658$). Crucially, and inconsistent with domain dichotomy, neither WD condition had higher familiarity than the BD condition.

Recollection rates were analysed using an ANOVA with factors of test condition (*recall-to-accept/recall-to-reject*) and pair type (*BD/WD-Name/WD-Image*). A main effect of pair type [$F(2, 56) = 4.50, p = .015$] reflected lower recollection for WD-Image (0.15) than BD (0.24; $t(28) = 2.24, p = .033$) or WD-Name (0.26; $t(28) = 2.70, p = .012$) pair types, but WD-Name and BD conditions did not differ ($p = .517$). A main effect of test condition [$F(1, 28) = 12.08, p = .002$] reflected

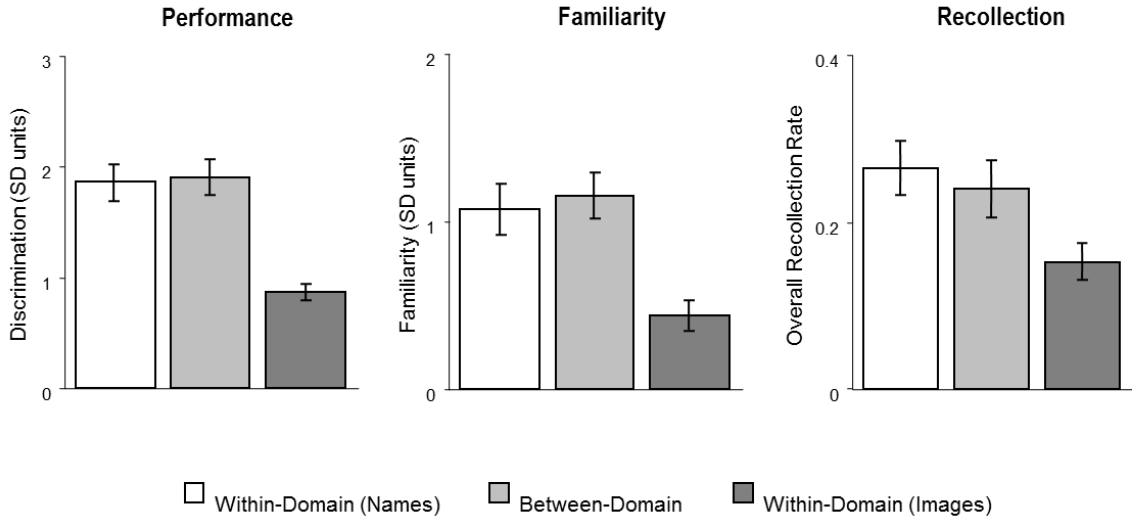


Figure 5.3: Mean discrimination, familiarity & recollection rates for Experiment 1. Shown separately for each condition and measured by fitting individual participant data to a DPSD model of associative recognition. Contrary to the domain dichotomy prediction, the BD condition exhibits just as much familiarity as the performance-matched WD-Name condition.

higher rates of recall-to-accept (0.27) than recall-to-reject (0.17); this did not interact with pair type ($p = .661$).

We assessed relationship-driven effects in two ways. First, items were matched across conditions by collapsing WD pairs together for each participant. Paired t-tests revealed stronger discrimination for BD than WD pairs (1.91 vs 1.22; $t(28) = 4.83, p = .001$; Cohen's d = 1.096); driven by greater familiarity (1.15 vs 0.75; $t(28) = 2.46, p = .021$; Cohen's d = 0.711), but not recollection (0.24 vs 0.21, $p = .337$). Second, we controlled for item effects by regressing discrimination, familiarity and recollection separately against factors of item (*two/one/zero names*) and relationship (*WD/BD*). Item type was significant for all three dependent variables: names led to better discrimination ($B = 0.991; p = .001$), familiarity ($B = 0.636; p = .001$) and recollection ($B = 0.112; p = .010$) than images. Relationship had a significant effect ($BD > WD$) on discrimination ($B = 0.547; p = .002$) and familiarity ($B = 0.392; p = .016$) but not recollection ($p = .391$). All reported effect sizes are unstandardized. Crucially, both methods reveal greater familiarity for BD than WD pairs, independent of item effects: the opposite pattern to that predicted by domain dichotomy.

5.2.3 Experiment 1 Discussion

Familiarity estimates from a DPSD model were significantly greater than zero for all pair types, consistent with a contribution of familiarity to associative recognition. Contrary to the prediction of domain dichotomy however, we observed greater familiarity for between-domain than within-domain pairs. Most importantly, this difference was present when controlling for stimulus class: analysis revealed independent effects of item type (stimulus class) and relationship (within/between-domain), and critically, when directly compared between-domain pairs were more familiar than within-domain pairs of the same items.

In examining the effect of relationship type we have used the familiarity estimate from the DPSD model. This provides a stronger test of the domain dichotomy prediction than familiarity as a proportion of overall recognition, which is as likely to reflect differences in recollection as familiarity. Nonetheless, others argue that a greater ratio of familiarity to accuracy for within-domain pairs may constitute evidence for domain dichotomy (Bastin et al., 2010). We therefore also compared proportional familiarity across conditions; this did not provide an alternative basis for supporting the domain dichotomy view (this analysis can be found as additional online material for Harlow et al., 2010). Given the results found in the preceding chapter, we also supplement this chapter with a re-analysis of the same data under the assumptions of a DPMSD model, which treats recollection as both graded and thresholded. The results from the DPMSD model are quite different to those implied by the DPSD model here and we explore them in more depth both in this chapter and the remainder of the thesis. Notably, however, these re-analysed data still do not support domain dichotomy. Finally, we also re-examined the data using an alternative unequal variance signal detection (UVSD) model (Green and Swets, 1966) to reinforce the conclusion that discrimination was greater for between-domain than within-domain pairs (1.83 vs 1.31; $t(28) = 3.83; p = .001$). In short, regardless of the approach taken to estimate memory processes, the ROC data are inconsistent with domain dichotomy theory.

5.3 Experiment 2

The DPSD model used to obtain process estimates in Experiment 1 is well suited to this purpose for two reasons: it generally gives a close fit to the ROC data, and it explicitly distinguishes between recollection and familiarity. Nonetheless, the model relies on a number of assumptions; consequently parameter estimates should be interpreted with caution, and preferably corroborated with other measures. In particular, the DPSD model assumes that familiarity and recollection are functionally independent. However if the processes are correlated, as in a redundancy view (Greve et al., 2010, Joordens and Merikle, 1993), both the DPSD model and the traditional remember-know paradigm would underestimate the true strength of familiarity for conditions eliciting high recollection.

To minimise the impact of this (unknown) statistical relationship on parameter estimates Mayes and colleagues (Mayes et al., 2007) suggest a modified remember-know procedure, whereby familiarity and recollection measures are obtained separately. Participants are trained to distinguish between familiarity and recollection (rather than the potentially misleading terms 'knowing' and 'remembering'). In a 'familiarity-only' procedure participants are asked not to actively recollect, but report recollection when it occurs. This measure of familiarity ought to be more reliable because the recollection rate is low and therefore the relationship between the two processes should have a small effect. In a 'recollection-only' procedure recall of some specific aspect of an original presentation is required for an old/new judgment, regardless of confidence. Making the distinction between strongly familiar and recollected trials explicit should result in more reliable estimates of recollection. Thus, Experiment 2 replicates Experiment 1, replacing ROC curves with the modified remember-know procedure.

5.3.1 Experiment 2 Methods

An additional 18 (10 female) participants (mean age 19.1, range 17-25) completed a modified remember-know procedure. The exclusion criteria, consent, ethics and payment rates were identical to those in Experiment 1. Each participant performed the familiarity-only task of the modified remember-know procedure for 6 consecutive blocks (2 of each condition) and the recollection-only

task for another 6 blocks; task order was counterbalanced across participants. In the familiarity-only task the *intact/rearranged* and confidence judgments at test were replaced with a single *familiar-intact/unfamiliar-rearranged/recollected* judgment. Participants responded intact or rearranged on the basis of familiarity only; when involuntary recollection occurred (of any aspect of an original study episode) they were required to respond 'recollected'.

In the recollection-only task participants made a single *recollected-intact/recollected-rearranged/no recollection* judgment. Here participants responded intact only if they recalled some aspect of the original study presentation, and rearranged if they recalled one of the items being paired with another at study. In the absence of explicit recollection they were required to respond 'no recollection', regardless of confidence. With the exception of these procedural differences Experiment 2 was identical to Experiment 1.

5.3.2 Experiment 2 Results

Non-recollected trials (of unknown accuracy) in the recollection-only experiment were assigned the (known) accuracy for non-recollected trials in the familiarity-only procedure, giving an overall accuracy for each participant and condition. Mean accuracy for each condition (BD = 0.81; WD-Name = 0.81; WD-Image = 0.66) did not reliably differ across Experiments 1 and 2 (BD: $p = .423$; WD-Name: $p = .441$; WD-Image: $p = .338$), suggesting that the change in retrieval task did not significantly alter performance. Familiarity was assessed by examining discrimination (false alarm corrected hits) in the familiarity-only procedure after discarding recollected trials[1]. Familiarity was lower for WD-Image (0.24) than BD pairs (0.38; $t(17) = 2.83, p = .011$), but WD-Name pairs (0.33) did not reliably differ from either WD-Image ($p = .385$) or BD ($p = .516$) pairs.

We similarly examined discrimination in the recollection-only procedure, after discarding non-recollected trials. Recollection-driven performance was poorer for WD-Image (0.42) than BD (0.70; $t(17) = 4.52, p = .001$), or WD-Name (0.71;

[1]This estimate of familiarity is accurate under an assumption of stochastic independence (recollected trials are, on average, no more or less familiar than non-recollected trials). We also assessed familiarity under the alternative statistical assumptions of redundancy (recollected trials are more familiar) and exclusion (recollected trials are less familiar). The results were qualitatively similar regardless of the assumption made.

$t(17) = 5.84, p = .001$) pairs, but BD and WD-Name pairs showed no difference ($p = .875$). Figure 5.4 summarises both familiarity and recollection-driven discrimination for each pair type.

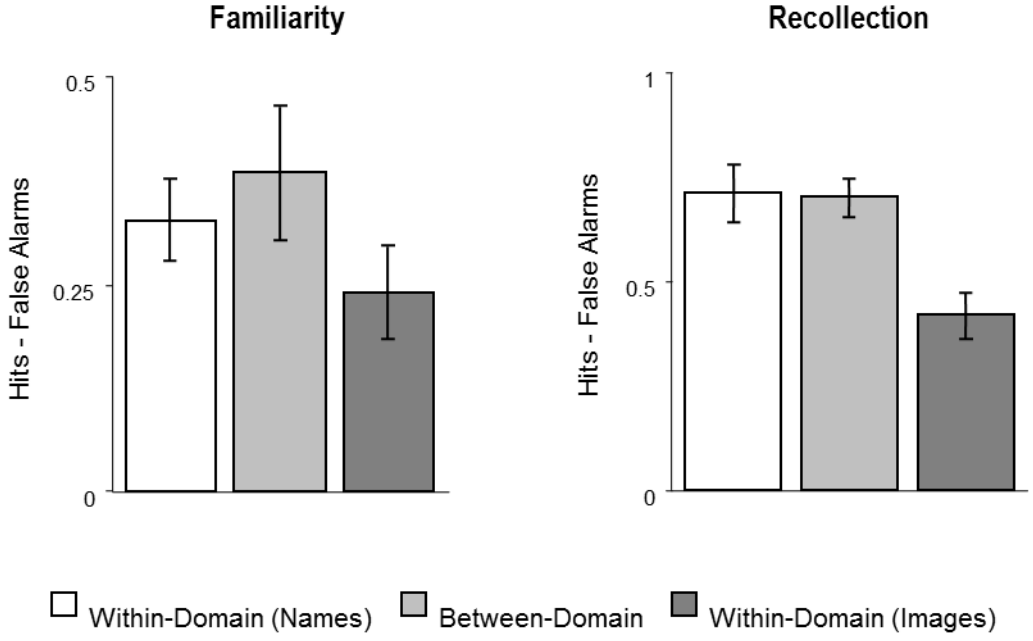


Figure 5.4: Mean familiarity- & recollection-driven discrimination for Experiment 2. Shown separately for each condition and measured using a modified remember-know procedure. The BD and WD-Name conditions did not differ for either familiarity or recollection, matching the pattern observed in Experiment 1.

We repeated the regression analysis from Experiment 1, with very similar results: names led to significantly better accuracy ($B = 0.125; p = .001$), familiarity ($B = 0.088; p = .028$) and recollection ($B = 0.293; p = .001$) than images, but BD pairs also led to significantly better accuracy ($B = 0.064; p = .031$) and familiarity ($B = 0.101; p = .022$) than WD pairs, with a marginal effect on recollection ($B = 0.136; p = .059$). When compared directly, BD pairs exhibited greater accuracy (0.81 vs 0.75; $t(17) = 3.30, p = .004$; Cohen's d = 0.679), recollection (0.70 vs 0.57; $t(17) = 2.17, p = .044$; Cohen's d = 0.516) and familiarity (0.38 vs 0.28; $t(17) = 2.13, p = .048$; Cohen's d = 0.549) than WD pairs of the same items.

5.3.3 Experiment 2 Discussion

The results from Experiment 2 closely match those from Experiment 1: familiarity appears to support performance in all three conditions, but in contrast to a domain dichotomy view the contribution was greater for between-domain than within-domain pairs. Of particular importance is the demonstration of phenomenological evidence for familiarity, given that familiarity estimates from the DPSD model rely upon an assumption that recollection is thresholded. If recollection is graded, the curvilinearity that is interpreted as reflecting familiarity could be accounted for by weaker recollection. While possible, this explanation seems to be inconsistent with above-chance performance in the familiarity-only procedure. In addition, participants all reported themselves well able to distinguish between familiar and recollected trials, both during the practice phase and at the end of the study. Thus, together our results suggest that performance is being supported by a process that both looks (Experiment 1), and feels (Experiment 2), like familiarity.

5.4 Discussion

The results presented here provide evidence that familiarity can contribute to the retrieval of novel associations. Our data suggest that familiarity supported performance in an associative recognition task, regardless of pair type (names, images, mixed pairs) or how performance was assessed (ROC analysis, modified RK procedure). As illustrated in Figure 5.5, however, familiarity was consistently greater for between-domain pairs. These results therefore present a fundamental challenge to domain dichotomy theory, raising questions about how familiarity should best be characterised and what role it plays in associative recognition.

The results of any study evidently rely to some extent on the stimuli used, and at present there is no precise definition of a domain to guide this choice. Perhaps, therefore, our particular stimuli simply do not give rise to overlapping representations as predicted. Data from neuroimaging may be important in this regard: future studies should demonstrate whether individual classes of stimuli are indeed represented in separate domains and whether item representations converge spatially (using fMRI) and temporally (using EEG). More broadly, in functional

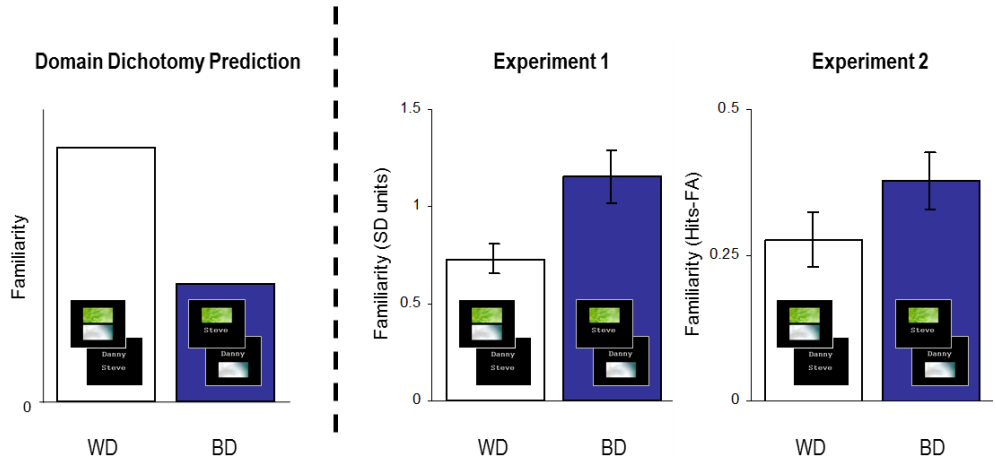


Figure 5.5: Summary of domain dichotomy prediction, with observed results from Experiments 1 & 2. Between-domain pairs elicited greater estimates of familiarity in both experiments, clearly contradicting the predictions of domain dichotomy theory.

terms, perhaps familiarity is not well characterised by the kinds of 'tuning' mechanisms and overlapping representations that are proposed by the models that motivate domain dichotomy theory (for further discussion see Greve et al., 2010).

Although our findings are clear, they stand in contrast with a study that claims support for domain dichotomy (Bastin et al., 2010), in which face-face pairs were shown to elicit more proportional familiarity than face-name pairs. These data are strikingly consistent with the lesion data reported by Mayes and colleagues, and the use of forced-choice procedures may account for some differences with our study. However, we have reservations about the strength of evidence they provide for domain dichotomy: face-name pairs actually gave rise to better associative recognition during pilot testing (familiarity was not reported) and face-face pairs were therefore presented for longer at study to equate performance. Given this manipulation it is possible that familiarity, like overall recognition, may have originally been matched or greater for face-name pairs - unfortunately the design of the experiment makes this impossible to determine. Even more importantly, between-domain pairs were not compared to within-domain pairs from both domains, making it mathematically impossible to disentangle (or characterise their result in terms of) item and relationship effects.

In our findings the relationship effect is demonstrably independent of item effects. Proponents of domain dichotomy might argue that the predicted effect is

still present, masked by a larger effect in the opposite direction - a possibility that is, of course, impossible to rule out. Thus here we focus on the key finding that between-domain pairs were recognised more easily than within-domain pairs: why might this be? In both experiments between-domain pairs elicited greater estimates of familiarity. This raises the possibility that they might be more robustly unitized than within-domain pairs, given that unitization has been implicated in familiarity for associations (Quamme et al., 2007, Rhodes and Donaldson, 2007; 2008). It is however circular to categorise stimuli as unitized (or not) based on differences in familiarity alone, emphasising the need for independent means of assessing unitization. Memory for unitized pairs might be more strongly impaired by manipulations that introduce perceptual differences between study and test (e.g., switching the positions of items, or presenting them separately), or recognition or perception of the individual components might be reduced following unitization (Mayes et al., 2007).

One aspect of the current findings is not predicted by unitization: in the second experiment between-domain pairs elicited higher levels of recollection compared to within-domain pairings of the same items. Given that unitization is primarily an account of familiarity, it is compatible with this change in recollection, but does not readily explain it. Instead, better memory for individual items might assist recollection and thereby support stronger associative recognition. For example, items might be more distinctive when presented as part of a between-domain pair, and therefore better recognised (Curran et al., 2002). Others have shown that increasing the number of similar (but not dissimilar) items in a list impairs memory, predicting poorer item recognition for within-domain conditions (Criss and Shiffrin, 2004). Intriguingly however, the same article suggested that associative recognition performance was dependent on the similarity of pairs rather than items, posing a challenge for a purely item-level explanation.

Finally, our data also demonstrate that the nature of the stimuli is important for remembering: names were generally better remembered than images. The relationship-driven difference we report here is, however, statistically independent of this item type effect. Interestingly, a previous study using faces and words (Criss and Shiffrin, 2004: Experiment 1, Group A) finds a similar effect of relationship, also independent of stimulus type effects. An important aim for future research will be to establish whether, in broader terms, relationship effects

are influenced by the nature of stimuli (Rhodes and Donaldson, 2007), or exist generally for certain types of association (e.g., within- or between-domain).

We began this study in search of evidence for domain dichotomy in one area where it has been notably lacking: psychological studies of normal subjects. While our results do not support domain dichotomy, they are consistent with a role for familiarity alongside recollection in associative recognition. Interestingly, they also suggest that the way items are combined might change the contribution of each process to retrieval - characterising these relationship effects remains an important goal for future research.

5.5 Supplementary analysis: Use of a mixture model

The chapter as presented thus far, with some minor differences, was published as (Harlow et al., 2010). Since that publication, we completed the experiments and analyses related in Chapter 4, concluding on that basis that recollection should be modelled as a graded and thresholded phenomenon, and that its all-or-none nature implied by the DPSD model is a simplification. Here we follow-up that conclusion with a practical question: does the choice of dual-process model have a significant effect on the conclusions drawn from our ROC data?

Specifically, it may be inappropriate to use a DPSD model to determine whether familiarity contributes significantly to associative recognition (i.e. the approach we took above). This is because when familiarity is estimated using a DPSD model, two assumptions are being made about recollection. First, recollection is thresholded (and familiarity is not); second, recollection is not graded. By contrast, the mixture model requires only the first assumption to be true in order to measure the presence of diagnostic familiarity in a given task. Since there is strong evidence that recollection is both thresholded (Chapter 4) and graded (Mickes et al., 2010, Slotnick, 2010), the first assumption, but not the second, is a reasonable one.

Thus instead we re-analysed the data using the alternative dual-process *mixture* signal detection (DPMSD) model and the related VRDP model, both described in Section 2.3.3. We also compared results obtained by fitting two simpler signal detection models, EVSD and UVSD, which do not distinguish between familiarity

and recollection. For both the VRDP and DPMSD models we compared fits both when familiarity was assumed to operate, and when it was assumed not to provide any information (i.e. the value of d'_F was set to 0).

5.5.1 Model fit statistics

Initial model fits were assessed using the chi-square goodness-of-fit statistic, after fitting each model to all 29 individual participants' data (Table 5.1). This comparison suggested that UVSD, EVSD and DPSD models should be rejected on the basis that their fit was relatively poor for the number of parameters used. We also calculated Akaike Information Criteria (AIC) for each model to allow the parsimony of each model to be compared, though see Section 3.2 where we discuss some reservations about the usefulness of AIC. The EVSD model provided the most parsimonious fit according to this measure, in contrast to the conclusion drawn from the chi-square statistics, followed by the VRDP model with d'_F set to 0 and the UVSD model.

We concluded from the model fits that the DPSD model provided a relatively inaccurate description of the data; the EVSD and UVSD models provided fits which were even less accurate but, according to AIC, more parsimonious (i.e. they were efficient at approximating the data). All four mixture models provided a more accurate account of the data than DPSD, and are differentiated in terms of their inclusion of the familiarity (d'_F) and recollection variance (ratio of s.d. of recollected to that of non-recollected trials, $v(R)$) parameters. We followed up this broad model comparison with a more focused test of the significance of each, using likelihood ratio tests. These indicated that fixing the variance of recollected and non-recollected trials to be equal did not significantly worsen the model fit ($\chi^2(29) = 36.3; p = .165$), except when familiarity was also included in the model, ($\chi^2(29) = 49.8; p = .009$). Including familiarity, however, did not improve the fit, regardless of whether the variance ratio was also included ($\chi^2(87) = 59.9; p = .988$) or not ($\chi^2(87) = 46.4; p = .999$). According to these analyses the best compromise between fit and parsimony was provided by the VRDP model with no familiarity.

Using this model, familiarity did not contribute significantly to performance. One-way repeated measures ANOVA revealed that recollection strength did vary

Model	Parameters	X^2	df	p	-LL	AIC
DPMSD	13	928	928	.489	11862	171
VRDP	12	993	957	.206	11887	162
DPMSD ($d'_F = 0$)	10	983	1015	.759	11892	57
VRDP ($d'_F = 0$)	9	1026	1044	.645	11910	35
DPSD	9	1120	1044	.050	11951	118
UVSD	6	1214	1131	.043	12006	53
EVSD	3	1339	1218	.008	12066	0

Table 5.1: Summary of model fits to the ROC data from Experiment 1. The number of parameters shown is the number of free memory-related parameters (9 criteria were also fit) per participant. X^2 denotes the total chi-square fit statistic summed across all participants, df similarly denotes the total degrees of freedom for the chi-square distribution and p denotes the resulting probability of the observed data under the null hypothesis (which is that the model does not fit the data). The negative log-likelihood of the data under the assumptions of each model is denoted by -LL, and summed across all 29 participants (lower numbers indicate better fits). AIC is the (relative) Akaike Information Criterion, a log-likelihood based statistic which penalises models according to the number of free parameters they contain (the AIC of the best fitting model is set to 0 for ease of comparison).

across conditions $[F(1.57, 43.89) = 10.01, p = .001]$ and that this was driven by lower strength to WD-Image (1.04) than WD-Name (1.34, $p = .004$) or BD pairs (1.40, $p = .001$), which did not differ from each other ($p = .339$). Recollection rates showed the same pattern when they were analysed using 2-way repeated measures ANOVA. A main effect of pair type $[F(2, 56) = 8.36, p = .001]$ was driven by lower recollection to WD-Image (0.42) than BD (0.60; $p = .004$) or WD-Name (0.61; $p = .001$) conditions, but WD-Name and BD conditions did not differ ($p = .791$). Intact pairs were only marginally more frequently recollected than rearranged pairs $[F(1, 28) = 3.665, p = .066]$ and did not interact with pair type ($p = .763$). When recollection was allowed to vary in strength, the advantage for between- over within-domain pairs was found to be driven by recollection and not familiarity.

5.5.2 Conclusion: The importance of model selection

Regardless of whether the variance of confidence ratings differed between recollected and non-recollected trials, the critical result here is that familiarity did not appear to contribute significantly to the associative recognition task. This is in stark contrast to the conclusion reached using a DPSD model above, a model which we have shown here fits the data less well, and which we know from Chapter 4 models recollection incorrectly. Proponents of the DPSD model acknowledge that it is not likely to be a perfect reflection of memory in all cases, but argue that in general the estimates it provides should be useful and informative given reasonable precautions (Yonelinas et al., 2010). The analyses here, in contrast, demonstrate that it is critical to use an accurate dual-process model to analyse ROC data.

Chapter 6

Component Recognition

In the previous study we found a consistent associative discrimination advantage for pairs of dissimilar (between-domain) items compared to pairs of similar (within-domain) items. That is, pairs comprising one name and one image were better recognised than when the same components were studied as name-name and image-image pairs. In this chapter we investigate the possibility that item-level memory differences may explain this effect. Specifically, we ask whether recognition or recall of component items differs across conditions and if so, what effect this might have on associative recognition.

6.1 Introduction

One reason to expect that component recognition may differ across conditions is that associative recognition appeared to rely heavily on recollection in the previous chapter. Namely, between-domain pairs exhibited significantly elevated recollection in Experiment 2, and when reanalysed using a more accurate dual-process model the same pattern was observed in Experiment 1. Discrimination of novel associations on the basis of recollection is likely to rely to some extent on successful recognition and recall of the associated components, for example as a cue for retrieval of a studied pair in recall-to-reject. Therefore, recollection differences between conditions in previous studies might reflect differences in the recognition or recall of individual components.

Pairs of names were easier to recognise than pairs of images. Importantly how-

ever, the between-domain advantage was apparently independent of this effect and resulted in greater performance than would be predicted by the type of components in the pair. A linear regression analysis and a direct comparison of between and within-domain pairs both suggested that differences in the type of components, and therefore their recognition, did not fully explain the advantage observed for between-domain pairs.

One interesting conclusion that might be drawn as a result is that the ability to later recognise a pair of components is determined by the general relationship between them, and not only their specific type. Such a conclusion can only be drawn, however, under the assumption that individual items are equally well-remembered whether they were part of a between-domain or within-domain pair. If in actual fact components of between-domain pairs are more easily recognised than the equivalent components of within-domain pairs, this could in turn lead to a corresponding advantage for between-domain pairs in associative recognition. Below we briefly outline two reasons why individual components might be expected to be better recognised when they are studied as part of a between-domain pair.

6.1.1 Material-specific list length effects

As noted in the discussion of the previous chapter, one way in which recognition might be impaired is by increasing the number of similar stimuli in a study list. For example, Criss and Shiffrin (2004) demonstrated that the similarity of stimuli, as well as the number of them, may be crucial in this respect. In other words, increasing the number of similar items in a list impairs memory for those items, but increasing the number of dissimilar items has little or no impact.

If a similar list length effect exists in our previous associative recognition studies, it might be expected to lead to improved recognition for components of between-domain pairs relative to those of within-domain pairs. This is because each study phase of n pairs contains 2n individual names or images in the within-domain conditions, but n names and n images in the between-domain condition. If adding a similar item to the list impairs memory more than adding a dissimilar item, components of between-domain pairs should be better recognised at test than those of within-domain pairs. If this is the case, better associative recognition of

between-domain pairs may partly or entirely reflect improved recognition of their components.

6.1.2 Presentation-level differences

List length effects are not the only possible source of variability in associative recognition performance. So far, we have referred to 'relationship-type' effects, meaning influences on associative recognition caused by the manner in which component items are combined, as opposed to any inherent properties of the components themselves. There is, however, at least one way in which the structure of between-domain pairs might lead to better-than-expected component recognition, but which might not be accurately described as a relationship-type effect. At study, pairs of items were presented on the screen for a limited period of time. If names and images differed in how much time was required to encode them effectively, this might lead to differences in component recognition for between compared to within-domain pairs.

Differences in the processing required by each stimulus category are likely given their differing semantic and perceptual properties, and therefore their underlying representations. For example, names are associated with some level of pre-experimental familiarity, whereas the images used are considerably more novel. As a result, names might be perceived relatively quickly and held, accurately, in working memory after the stimulus presentation has ended. Images on the other hand may be harder to store accurately in working memory and therefore benefit from being attended to on screen for longer. By this view, during a between-domain presentation a participant might spend the bulk of the available time encoding the novel image and still be able to encode the name after the presentation has ended. This strategy should lead to better recognition of images when they are encoded as part of a between-domain condition, but no such advantage (and possibly even a disadvantage) for names. In contrast to the list length effects described above, presentation-level differences in component recognition should not be affected by intermixing conditions in each block, since the differences are produced at each individual presentation and not across the study list as a whole.

6.1.3 Testing component and associative recognition together

To investigate whether the component-level effects described above occur in our experimental design, and if so how they might influence associative recognition we tested both item (studied vs unstudied components) and associative (intact vs rearranged pairs) discrimination together within the same experiment. We also included a between-subjects manipulation: half of participants studied each condition in separate blocks as for previous chapters, and the remaining half studied intermixed blocks comprising all three stimulus conditions (and therefore equal number of names and images in total). This experimental design allows us to address several questions. Firstly, we can test directly whether component recognition varies across conditions. In particular we examine whether components are better recognised when they are studied as part of a between-domain pair, and if so whether this is a general effect for both item types or is specific to names or images. Secondly, we can test where this difference is produced. If it is a result of material-specific list length effects, between-domain components should be better recognised when the conditions are kept in separate blocks, but not when the conditions are intermixed. Alternatively, if component recognition is determined at the level of individual pairs, intermixing conditions should have no effect on differences across conditions. Thirdly, by testing item and associative recognition within participants, we can examine how the two are related. For example, does associative recognition performance correlate linearly with component recognition, or is the relationship more complex? Does the relationship differ across pair types, suggesting condition-related differences in how associations are encoded or retrieved? Most importantly perhaps, if component recognition differences exist, is it plausible that they can explain the associative recognition advantage for between-domain pairs?

6.2 Methods

Two groups of participants studied pairs of items; half doing so in blocks of the same pair type ('separate') and half of participants studying all three pair types in each block ('intermixed'). At test, all participants were presented with two intermixed tasks: discriminating between intact and rearranged pairs of the

same items, and discriminating between old and new individual components. On each trial, participants made either an item or associative recognition judgment as appropriate and then rated their confidence from 1-5 after each decision, allowing memory performance to be compared for the two tasks using the discrimination estimate d_a, and estimates of familiarity and recollection to be drawn by reference to memory strength models.

6.2.1 Participants

A total of 36 right-handed, native English speakers were randomly assigned to either the 'separate' or 'intermixed' group for the Experiment and all data sets were included in the final analysis. One group of 18 participants (12 female; mean age 21.1, range 18-27) studied the three stimulus conditions (*WD-name*/ *WD-image*/ *BD*) in separate blocks, exactly as in Chapter 5. The remaining 18 participants (12 female; mean age 19.2, range 17-30) studied blocks comprising all three conditions intermixed. Importantly, in all other ways the experiment was identical for both groups.

6.2.2 Stimuli

Stimuli are described in Chapter 3 (see Figure 5.1). The three stimulus conditions were either presented in separate study-test blocks ($N = 18$), as in the previous chapter, or were intermixed in equal proportions ($N = 18$).

On each trial at test, participants were shown either a pair of items in the same positions as at study, or a single item presented in the centre of the screen. For pairs of items, an associative recognition judgment (*intact / rearranged*) was made, similar to the previous study. All of these pairs comprised items which had been seen exactly once during the study phase. In contrast, half of the single items presented were previously-studied (as part of a pair) and the remaining half were unstudied, therefore an item recognition judgment (*old/new*) was made for these items (Figure 6.1). It was explained clearly to participants before commencing the experiment that both item and associative recognition would be tested.

Names were screened for length (4-8 letters) and images were selected from a pool of 575 images, rated for abstractness by a separate group of participants

(see Section 3.1.2). The selected images were rated "abstract" 74% of the time. In total 432 names and 432 images were used.

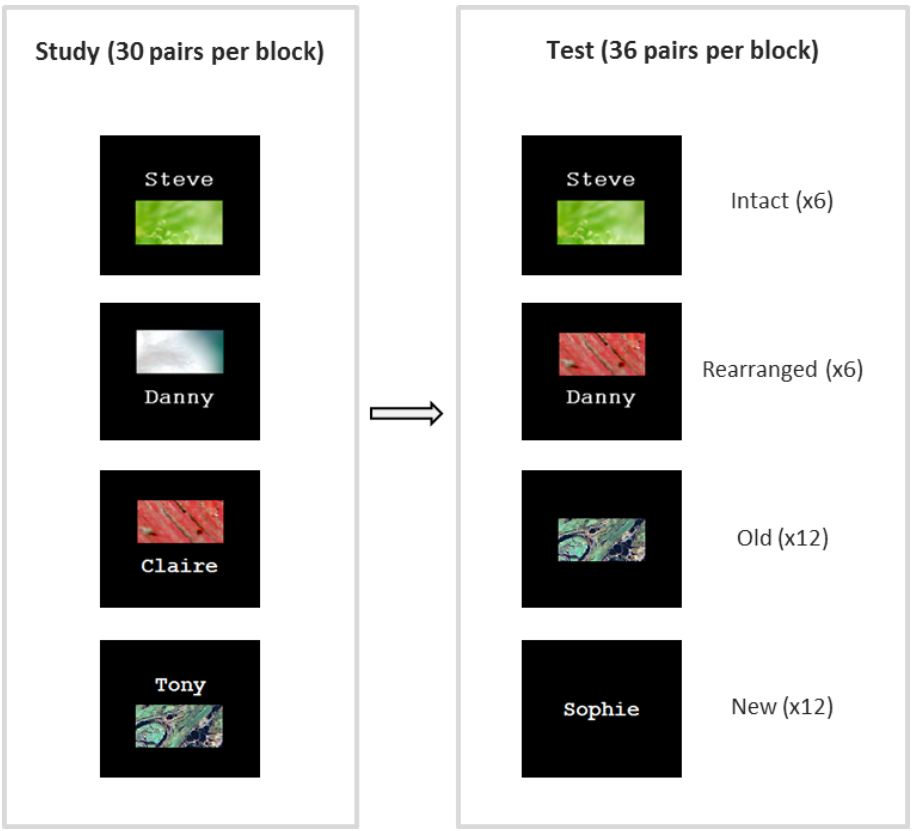


Figure 6.1: The combined item/associative recognition task. Each study phase comprised 30 stimulus pairs and each test phase comprised 36 presentations of four randomly intermixed types. *Intact* and *Rearranged* pairs were constructed using components from the preceding study phase, each of which was presented in the same position (above or below fixation) as at study. *Old* and *New* items were individual components from the preceding study phase or unstudied items presented in the centre of the screen. Participants made an intact/rearranged judgment to test pairs or an old/new judgment to items, followed in each case by a confidence rating from 1-5.

6.2.3 Procedure

Detailed aspects of the experimental procedure not included here can be found in Chapter 3. The experiment was divided into 12 blocks, and each block was further subdivided into a study phase of 30 trials followed by a test phase of 36 trials. Each study trial contributed to exactly one trial at test: 6 pairs were later

shown intact, single items from 12 pairs were recombined to form 6 rearranged test pairs, and single items from the remaining 12 study pairs were shown as old items at test (Figure 6.1).

The study and test procedures are illustrated in Figure 6.2. Each study trial was preceded by a 1000ms fixation cross, and consisted of a pair of items presented for 2000ms above and below central fixation. After each study presentation participants were required to indicate on a scale from 1-5 how well the two items went together; this response initiated the beginning of the next study trial.

At test, following a 1000ms fixation cross, participants were presented with either a single item for 800ms at central fixation (item recognition presentation) or a pair of items for 1000ms above and below central fixation (associative recognition presentation). Following each item recognition presentation participants judged the item to be old or new, where old items were those which had appeared as part of a pair during the preceding study phase, while new items had not been previously shown at all. Associative recognition presentations were judged intact or rearranged, exactly as in the previous chapter. Both types of decision were followed by a 500ms blank screen, after which participants indicated how confident they were that they were correct, again using a scale of 1-5. This confidence response initiated the beginning of the next test trial. Associative and item recognition trials were randomly intermixed at test and in total each participant completed 24 intact, 24 rearranged, 48 old and 48 new trials per stimulus condition.

For the 18 participants assigned to the 'separate' group, each block comprised pairs and items from a single stimulus condition, 4 blocks for each condition, as in the previous chapter. For the 18 participants assigned to the 'intermixed' group, all 12 blocks comprised an equal number of study trials (10) for each stimulus condition, as well as an equal number of intact (2), rearranged (2), old (4) and new (4) test trials. Thus, for both the intermixed group and the between-domain condition of the separate group, exactly half of the items in each study and test phase were names. For the name-name and image-image conditions of the separate group, all of the items were names or images respectively.

At both study and test the mapping of left and right buttons to (*old*/*new*) and (1-5) responses was identically counterbalanced across blocks of 4 participants for

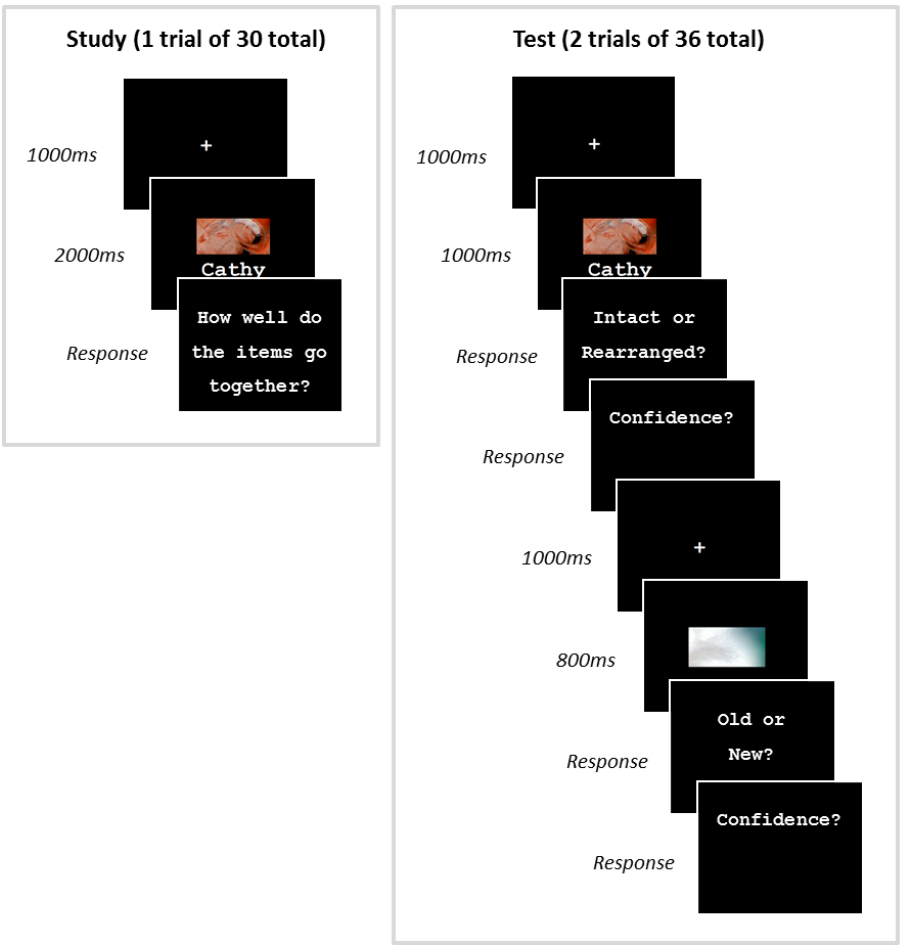


Figure 6.2: Study and test procedures. Participants studied 30 stimulus pairs for 2000ms each, judging how well the components of each pair appeared to go together. At test, 12 associative recognition trials consisted of pairs of previously-studied items (6 intact, 6 rearranged) presented for 1000ms in the same screen positions as at study. Participants judged each pair to be intact or rearranged. Twenty-four item recognition trials consisted of individual components (12 old, 12 new) presented for 800ms at central fixation, to which participants made an old/new judgment. The two types of test presentation were randomly intermixed and after each one participants rated their confidence on a scale of 1-5. Every screen was followed by a blank (black) screen for 500ms (not shown in the figure), with the exception of the fixation cross, which was followed by a 100ms blank screen.

each group, intact and old responses were made using the same button. The stimulus condition (3: *WD*/*WD*/*BD*) and test condition (6: *intact*/*rearranged*/*not shown*/*old*/*new*/*not shown*) of each component was fully counterbalanced within and across groups. On average the procedure took 2 hours to complete, including a practice block and debriefing.

6.3 Results

We assessed associative recognition using d_a, calculated directly from confidence ratings. Performance is summarised in Figure 6.3. Visual inspection suggests that discrimination was elevated for between- compared to within-domain pairs and also that name-name pairs were better recognised than image-image pairs, but intermixing conditions within blocks had no clear effect on associative recognition. Item recognition performance is summarised in Figure 6.4, which shows that images are better recognised when studied as part of a between-domain pair than as part of a within-domain pair. Performance across the two tasks is analysed in detail below.

6.3.1 Associative recognition performance

We first tested whether each experimental factor affected associative discrimination, using linear regression. We included factors of item type (0 *names*; 1 *name*; 2 *names*), relationship type (*within-domain*; *between-domain*) and group (*separated*; *intermixed*). Replicating findings from the previous chapter, we found a significant effect on performance from both item type ($B = 0.721; p < .001$) and relationship type ($B = 0.509; p = .007$), reflecting improved discrimination for pairs comprising names rather than images, and better discrimination for between- than within-domain pairs. In contrast, the effect of the between-subjects group factor (*separated*; *intermixed*) was not significant, ($B = -0.063; p = .737$), confirming that intermixing conditions within blocks had no measurable effect on associative discrimination. Table 6.1 summarises the results of the regression analysis.

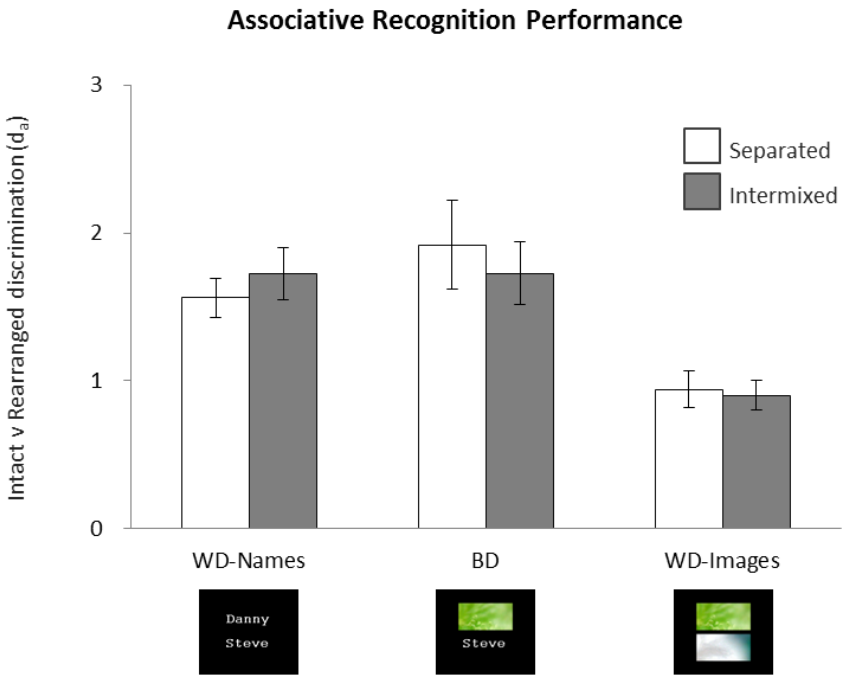


Figure 6.3: Summary of associative recognition performance. Mean discrimination d_a of intact from rearranged pairs is shown as a function of stimulus condition (*WD-names*, *BD*, *WD-images*) and group (stimulus conditions *intermixed* or *separated* in each study-test block). Replicating results from previous experiments, discrimination was significantly affected by item type (names lead to better discrimination) but also relationship (between-domain lead to better discrimination). Intermixing or separating stimulus conditions in each block had no significant effect on associative recognition performance.

6.3.2 Item recognition performance

Performance on the item recognition task, as measured by old/new discrimination, is summarised in Figure 6.4. We investigated how old/new discrimination varied using a repeated measures ANOVA with within-subject factors of item type (*image*; *name*) and relationship type (*within-domain*; *between-domain*) and a between-subjects factor of group (*separated*; *intermixed*).

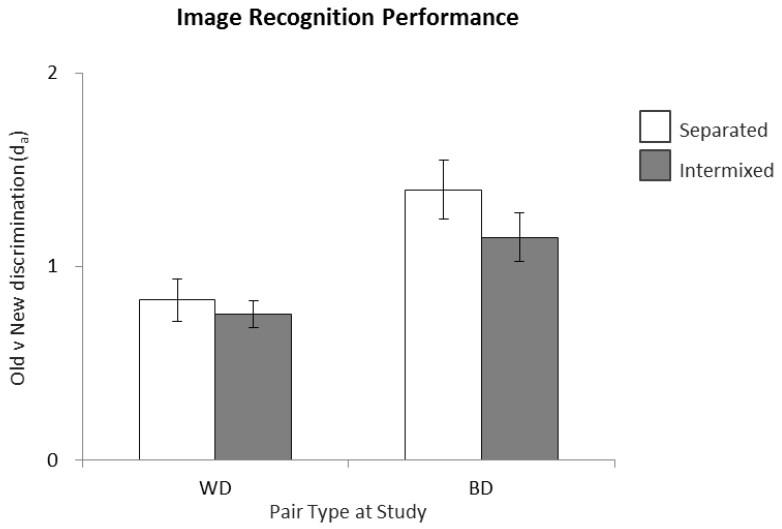
Not unexpectedly, item type had a significant effect on old/new discrimination $[F(1,34) = 29.30, p < .001]$, such that names were easier to recognise than images. Relationship type also had an effect in that items which had originally been studied as part of a between-domain pair were more easily recognised than those studied in within-domain pairs $[F(1,34) = 14.96, p < .001]$. This relationship effect

Parameter	Direction	B	S.E.	β	t	p
(Constant)		0.892	0.161		5.551	.000
Item type	Names > Images	0.721	0.185	0.340	3.887	.000
Relationship type	BD > WD	0.509	0.185	0.277	2.744	.007
Group	Intermixed < Separate	-0.063	0.185	-0.034	-0.337	.737

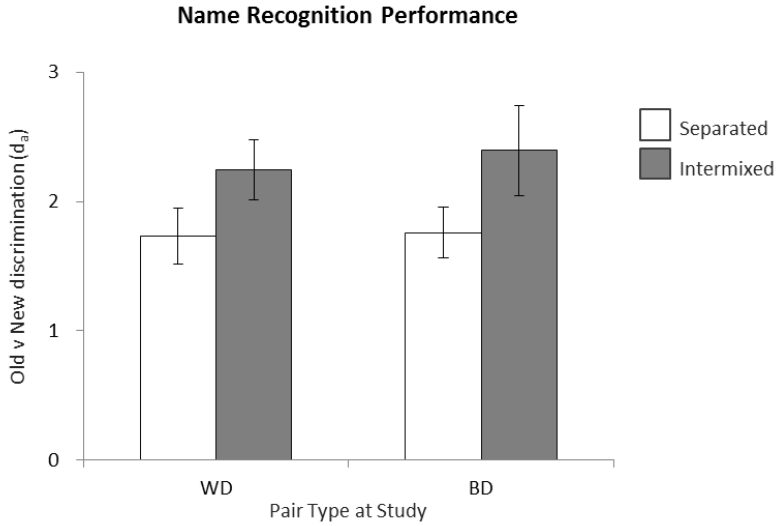
Table 6.1: Linear regression factors contributing to associative discrimination. Item type and relationship type both predict discrimination of intact from rearranged pairs. Pairs containing names were more easily discriminated than those containing images, but the relationship between items had an additional effect on discrimination: performance was better for between- than within-domain pairs.

interacted with item type [$F(1,34) = 6.66, p = .014$], reflecting the fact that images were more sensitive than names to relationship type at study. In fact, follow-up paired t-tests revealed that while images were indeed much better recognised when studied as part of a between-domain pair [$t(35) = 5.78, p < .001$], Figure 6.4(a), recognition for names was insensitive to relationship type at encoding [$t(35) = 0.72, p = .479$], Figure 6.4(b).

In contrast, the effect of intermixing conditions within blocks was limited. For old/new discrimination, just as for associative discrimination, the main effect of group on performance was not significant [$F(1,34) = 1.29, p = .264$]. While a marginal interaction between group and item type was found [$F(1,34) = 3.94, p = .055$], we interpret this result with caution. Taken together with the absence of a main effect of group, it would require that intermixing conditions within each block should improve recognition of images, but actually impair recognition for names. If decreasing material-specific list length does improve performance however, it should do so for both types of item, and so the pattern in the data is inconsistent with the effect predicted. In conclusion, components of between-domain pairs did indeed show improved recognition over those of within-domain pairs, but only for images, and intermixing conditions at study neither improved nor impaired component recognition.



(a) Image recognition as a function of encoding condition and group.



(b) Name recognition as a function of encoding condition and group.

Figure 6.4: Summary of item recognition performance. Mean discrimination d_a of old from new items is shown as a function of encoding condition (*WD*, *BD*) and group (*intermixed* or *separated*) for (a) images and (b) names. Studied names were more easily discriminated from new names than images were. However, images were more easily recognised when studied as part of a between-domain pair than when they were studied as part of an image-image pair, leading to better recognition overall of components from between-domain pairs than would be predicted solely by item type. Intermixing or separating stimulus conditions in each block had no significant effect on item recognition performance.

6.3.3 How does component recognition predict associative discrimination?

The results above indicate that images studied as part of a between-domain pair are better recognised than those studied as part of a within-domain pair. To what extent does this advantage for component recognition explain the between-domain advantage for associative recognition? To address this issue we repeated the regression analysis from 6.3.1, this time controlling explicitly for differences in component recognition by including an additional factor of old/new discrimination (d_a for the corresponding item recognition task for every participant and condition). If component recognition differences alone are sufficient to explain the associative recognition advantage for between-domain pairs, the factor of relationship type should no longer be significant when old/new discrimination is also included as a factor.

Parameter	Direction	B	S.E.	β	t	p
(Constant)		0.466	0.154		3.034	.003
Old/New d'	High > Low	0.624	0.100	0.590	6.231	.000
Item type	Names < Images	-0.027	0.199	-0.013	-0.135	.893
Relationship type	BD > WD	0.519	0.159	0.283	3.265	.001
Group	Intermixed < Separate	-0.073	0.160	-0.040	-0.458	.648

Table 6.2: Independent linear effects of old/new discrimination and relationship type on associative discrimination. Here condition-specific differences in component recognition are explicitly taken into account by including old/new discrimination as a factor. Item type no longer significantly predicts associative discrimination, indicating that it does so only via changes in component recognition. In contrast, relationship type is still a highly significant predictor of associative recognition performance: improved recognition for between-domain pairs is due at least partly to some factor beyond the recognition of their components.

The regression analysis is summarised in Table 6.2. The between-subjects factor of group (conditions either *intermixed* or *separated* in each block) was again not a significant predictor of associative discrimination ($B = -0.073; p = .648$). Since

intermixing conditions within blocks had no reliable effect on either associative or item recognition, remaining analyses in this chapter are simplified by collapsing the data across the factor of group, resulting in one larger study cohort of 36 participants.

Old/new discrimination, in contrast, was a strongly significant factor ($B = 0.624; p < .001$), with better item discrimination leading to better associative discrimination. Item type (0 *names*, 1 *name* or 2 *names* per pair) was no longer predictive of associative discrimination performance after old/new discrimination was included ($B = -0.027; p = .893$), indicating that the item-level effect was fully explained by component recognition. Crucially, however, relationship type remained highly significant ($B = 0.519; p = .001$), highlighting that a relationship-level effect existed independently from any effects of old/new discrimination. After removing the non-significant factors of item type and group, old/new discrimination ($B = 0.612; p < .001$) and relationship type ($B = 0.483; p = .001$) were still both strongly significant predictors of associative recognition performance. Between-domain pairs were better discriminated than within-domain pairs, even after controlling for differences in recognition of their components.

The results of the linear regression analysis suggest that the relationship between items has a significant effect on their subsequent associative recognition, and that this effect is independent of how well the individual items can be recognised. Importantly, however, this conclusion rests on the assumption that the dependent factor of associative discrimination $d_a(associative)$, is linearly related to the independent factors of old/new discrimination $d_a(item)$. If, in fact, the two are nonlinearly related, and one condition has larger values of the independent factor than other conditions, this could manifest as a spurious but significant condition-specific factor in the linear regression analysis. In other words the relationship effect found above could, under these circumstances, simply result from greater old/new discrimination in the between-domain than within-domain conditions combined with a nonlinear relationship between old/new and associative discrimination.

To guard against the possibility that old/new discrimination was actually a non-linear predictor of associative discrimination, we checked that the two were linearly related. Figure 6.5 plots associative versus old/new discrimination for each participant and condition, revealing an approximately linear relationship

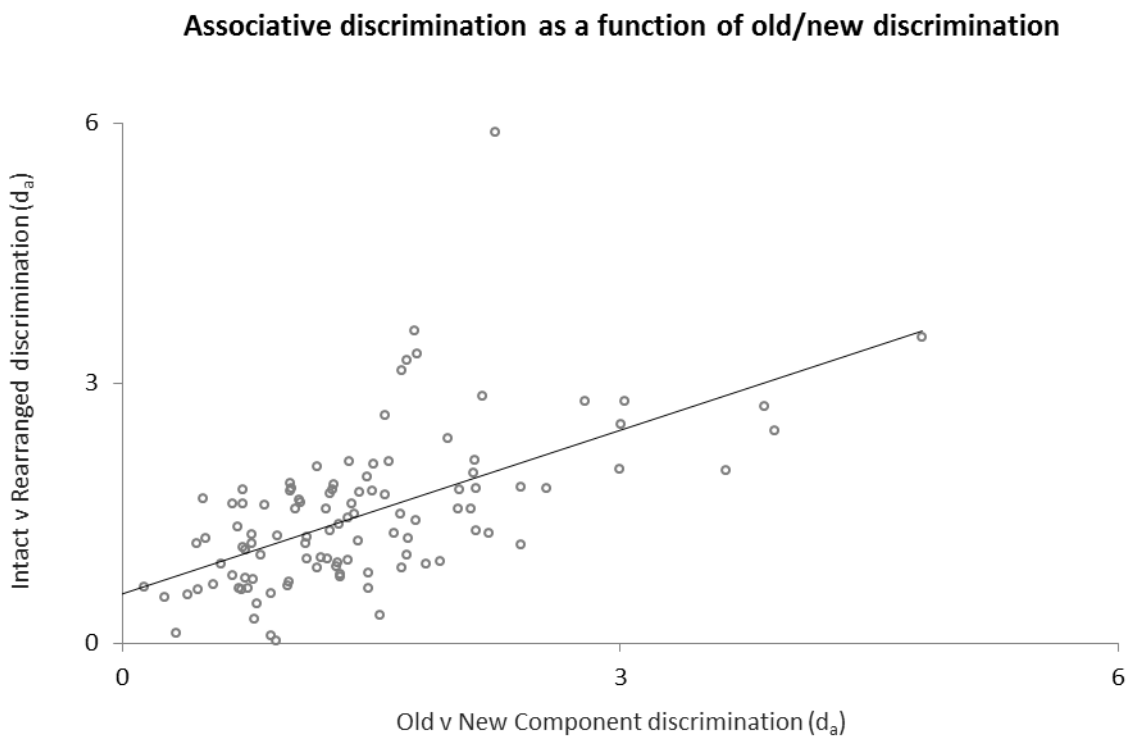


Figure 6.5: Associative discrimination as a function of old/new component discrimination. Discrimination of intact from rearranged pairs is positively correlated with and approximately linearly related to recognition for the individual components of each pair. This relationship justifies the use of a linear regression analysis which revealed that between-domain pairs are more easily recognised than within-domain pairs, independent of component recognition.

between the two; visually, no strong nonlinearities are apparent. We quantified this using a regression analysis including both linear and quadratic factors of old/new discrimination as independent variables. The linear factor was significant ($B = 0.909; p < .001$) but the quadratic factor was not ($B = -0.071; p = .227$), indicating a broadly linear relationship. The results of this regression are summarised in Table 6.3.

Since old/new and associative discrimination are approximately linearly related, the additional significant effect of relationship on associative discrimination is likely to represent a real and independent phenomenon. Nonetheless, to be certain that a small nonlinearity did not interact with elevated or more variable component recognition for between-domain pairs, we also examined the distribution of old/new discrimination across conditions. As Figure 6.6 reveals, however,

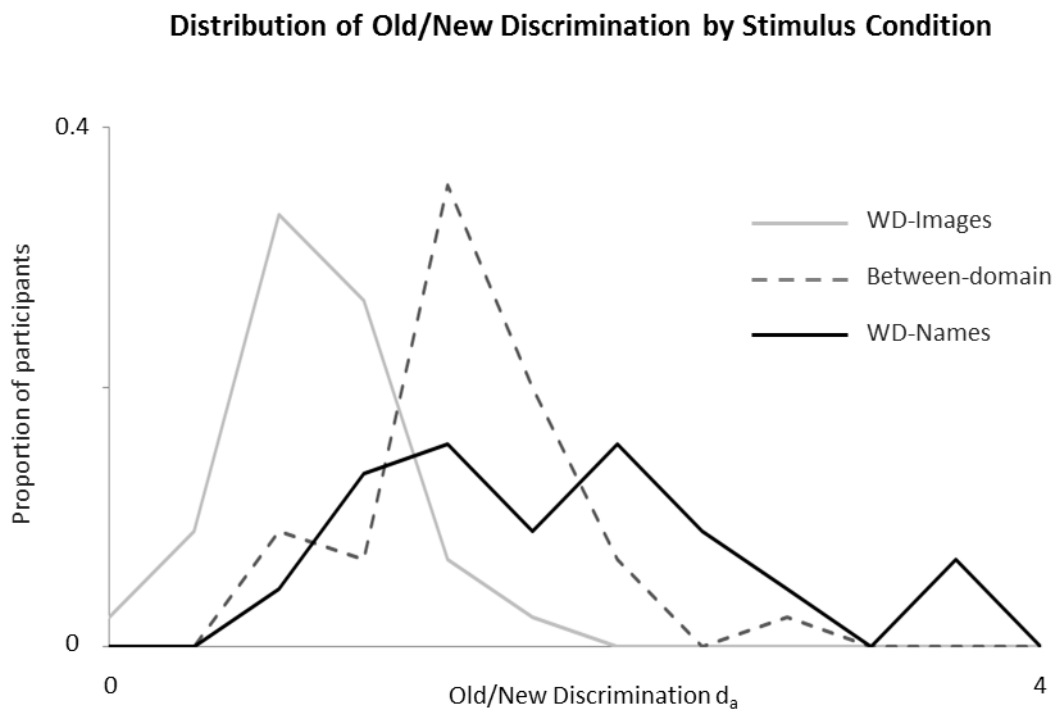


Figure 6.6: Distribution of old/new discrimination across participants. Mean recognition of the components of between-domain pairs is lower than for name-name components and higher than for image-image components, and no more variable than either. To facilitate direct comparison of the three conditions, area plots are used, rather than histograms, which is the normal practice. An area plot carries equivalent information as a histogram, with the main difference being that the lines interpolate between the counts at each bin. Each vertex corresponds to one observed data point: the height (i.e. frequency) of the corresponding bin.

between-domain pairs were not associated with either elevated or more variable component recognition; they had a distribution of old/new discrimination values broadly similar to those for name-name or image-image pairs, with if anything fewer extreme values than for within-domain pairs. In fact, the greatest differences in old/new discrimination values are between name-name and image-image pairs. Therefore, even if subtle nonlinearities did exist in the relationship between old/new and associative discriminations, these should manifest most strongly as item-type effects dissociating between name-name and image-image pairs. No such item-type effects exist. We can therefore be confident that the relationship-type effect is not an artefact of the analysis and genuinely reflects some advantage for between-domain pairs which is independent of component recognition.

Parameter	Direction	B	S.E.	β	t	p
(Constant)		0.362	0.220		1.650	.102
Old/New d' (Linear)	High > Low	0.909	0.246	0.859	3.689	.000
Old/New d' (Quadratic)	High < Low	-0.071	0.058	-0.283	-1.216	.227

Table 6.3: Linear and quadratic factors of old/new item recognition predicting associative discrimination. We regressed the associative discrimination for different participants and conditions against linear and quadratic factors of each corresponding component discrimination. The quadratic factor was not significant, indicating a broadly linear relationship between component and associative recognition.

6.3.4 Factor analysis of performance across tasks

Having ruled out the possibility that differences in associative recognition simply reflect differences in component recognition, we next consider the relationship between conditions in more detail. To what extent might the different tasks and conditions in this study rely on overlapping processes, and to what extent are they independent from eachother? To begin to address this question we reduced the dimensionality of discrimination scores using principal component analysis. The interpretation of factors obtained in this manner is not straightforward (see Appendix A for a discussion), and it is useful to clarify *a priori* expectations about the data. Firstly, two major sources of variance might be expected to give rise to large factors: item type (names or images) and task type (item or associative recognition). The order in which these are separated by the model might give some insight into whether differences in material or process are a more important determinant of memory performance. In contrast, if performance on between-domain pairs is uniquely supplemented by additional resources this should not be represented by a large factor, since between-domain pairs constitute only 14% of the trials at test. Instead therefore, the factor analysis might allow us to identify those conditions which are well-explained by large, interpretable factors, and those with considerable remaining unexplained variance which might suggest

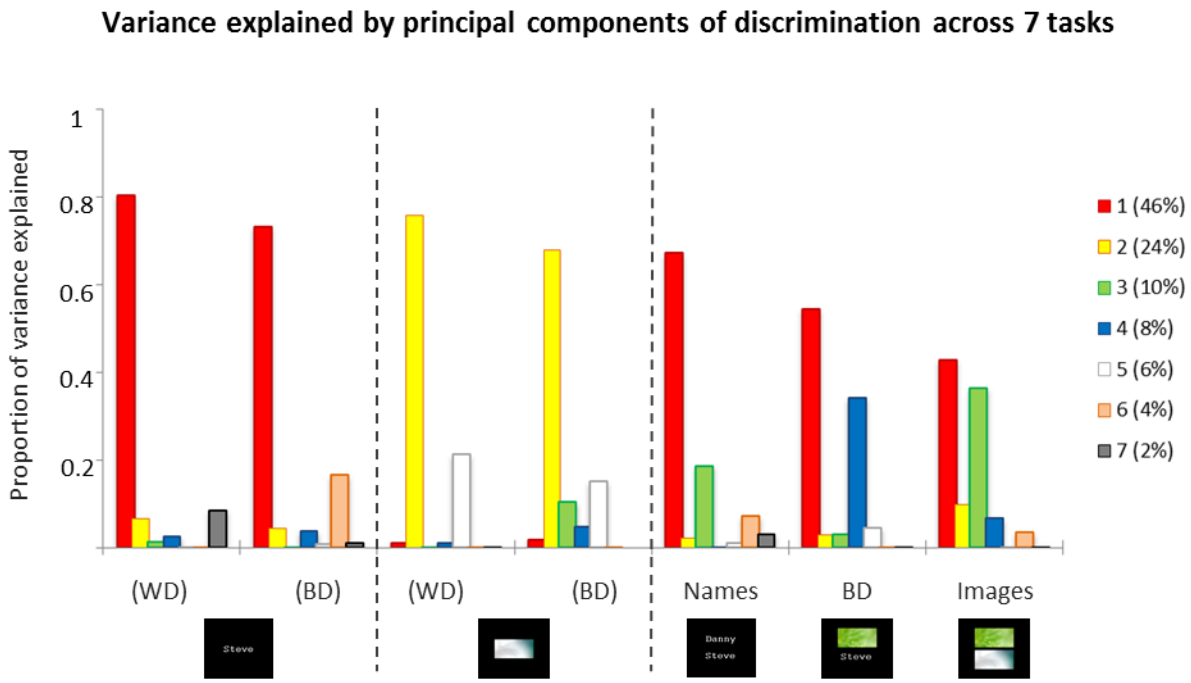


Figure 6.7: Variance in task performance explained by principal components. The proportion of variance in each of the 7 tasks which is explained by each factor is represented by the height of the corresponding bar. The first two factors separate the tasks into two intra-related groups: The largest factor accounts for between 40% and 80% of the variance in all three associative recognition tasks and item recognition for names, but virtually none of the variance in item recognition for images, which is primarily and uniquely accounted for by the second largest component.

different or less consistent processing for these conditions.

The proportion of variance accounted for by each extracted component is illustrated in Figure 6.7. The first component accounts for a considerable proportion of the variance in discrimination scores for individual names (WD 73.4%; BD 80.4%), name-name (67.2%), between-domain (54.6%) and image-image pairs (42.9%). In contrast, this first component has virtually no relationship with the ability to discriminate old from new images, explaining less than 2% of the variance in each case. Instead a second component, by definition of the extraction process uncorrelated with the first, accounts for over two thirds of the variance in discrimination of old from new images (WD 75.8%, BD 67.8%), but relatively little of that for names (WD 6.5%; BD 4.4%) or for associative discrimination (name-name 2.2%, BD 2.9%, image-image 9.9%). The largest source of within-

participant variance in the data seemingly resulted from differences in item type, with associative discrimination performance more closely related to name than image recognition.

Given a hypothesis about the number of factors in the model, a solution can be rotated in such a way that it maximises the variance of the squared factor loadings for each measured variable (VARIMAX). The result is to minimise the number of variables with high loadings on each factor, making each factor theoretically more interpretable by relating it more clearly to fewer variables. An alternative approach is to rotate the solution in such a way that the number of factors with high loadings is minimised for each variable (QUARTIMAX). We used VARIMAX and QUARTIMAX rotations on the largest 2, 3 and 4 components of the data to examine how different tasks might be explained by common factors, while avoiding making any strong assumptions about the number of factors in the model. The results for both rotations are similar to those for the unrotated solution shown in Figure 6.7, and are summarised below. Full results are included in Tables A.1 to A.3, and are included in the appendix on pages 281 to 283.

Recognition of stimuli consisting solely of names (individual names and name-name pairs) is predominantly explained by the same factor regardless of the number of components extracted or the rotation used, consistent with the view that these tasks all rely on closely overlapping underlying resources. Recognition of individual images is also consistently explained by a single factor. Image-image and between-domain discrimination are less well-explained by the two largest factors in the model, though if anything both rely more strongly on the first, name-related factor than the second, image-associated factor. Discrimination for image-image and between-domain pairs are accounted for by including a third and fourth factor respectively, however the proportion of variance accounted for by these two factors is limited, making it difficult to conclude from the factor analysis alone whether either represents a psychologically meaningful phenomenon. The overall pattern of results from the factor analysis, however, is consistent with the engagement of separate neural resources for recognition of image-image or between-domain pairs, compared to recognition of individual components or name-name pairs. In the next two sections we attempt to characterise the processes supporting the recognition of each stimulus type more concretely, firstly by examining the correlations between component and associative recognition and

secondly by comparing estimated contributions of familiarity and recollection for each condition.

6.3.5 Associative discrimination correlates with name but not image recognition

The factor analysis revealed two large factors which corresponded closely to the recognition of names or images respectively. Surprisingly, however, associative recognition performance appeared to be explained strongly by the name-related factor, but not by the image-related factor, regardless of the stimulus condition. To quantify this pattern statistically we examined the Pearson correlations between intact/rearranged discrimination for each stimulus condition with old/new discrimination of either names or images, collapsed across relationship type at study (Figure 6.8). Associative discrimination for all three pair types correlated significantly with old/new discrimination of names (name-name pairs, $R = .777, p < .001$; BD pairs, $R = .484, p = .003$; image-image-pairs, $R = .469, p = .004$). By contrast, none correlated significantly with old/new discrimination of images (name-name pairs, $R = .041, p = .811$; BD pairs, $R = .170, p = .322$; image-image-pairs, $R = .256, p = .132$). Results and p-values were qualitatively similar for Spearman's rho and Kendall's tau correlations.

Discrimination for all three associative recognition conditions correlated significantly with the ability to recognise names, but not images. This result corroborates, and provides statistical backing for, the suggestion from the factor analysis that closely overlapping processes may contribute to the recognition of pairs and individual names, but that a different combination of processes support the recognition of individual images. Under a dual-process view, one might hypothesise that differences in the relative contribution of familiarity and recollection across conditions might give rise to such a pattern; specifically that individual images were recognised primarily on the basis of familiarity, whereas associative recognition and name recognition relied more heavily upon recollection. In the following section we test this hypothesis directly by estimating the contributions of familiarity and recollection to each condition.

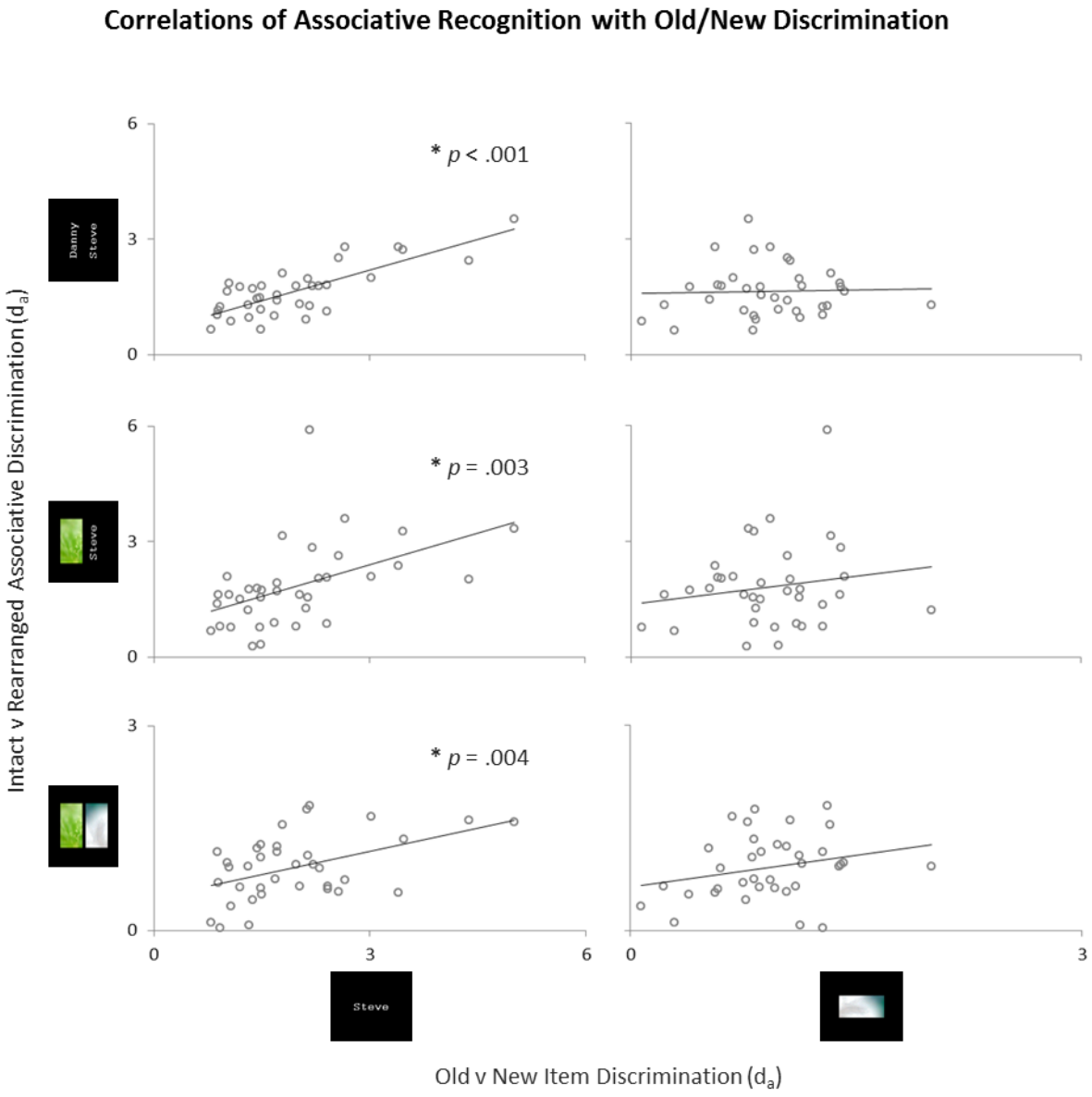


Figure 6.8: Correlations of associative recognition with old/new recognition. Associative recognition performance correlates with recognition of individual names, but not images, even for pairs consisting only of images.

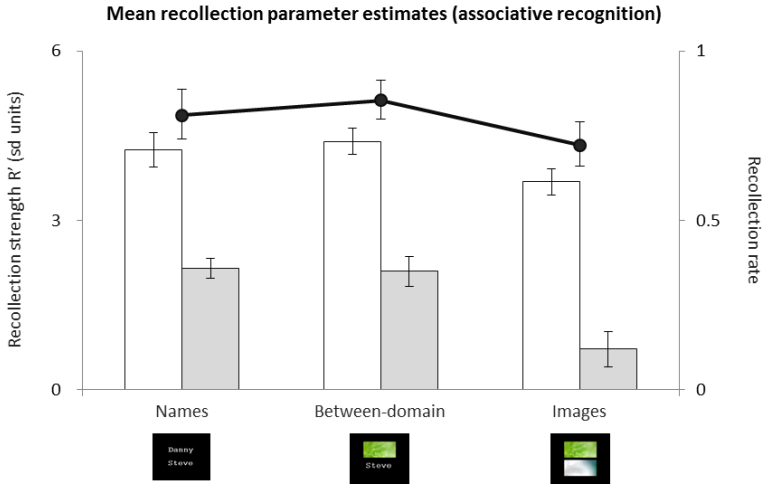
6.3.6 Familiarity and recollection estimates

To extract estimates of the contributions of recollection and familarity to recognition judgments we fit each participant's data to the dual-process mixture signal detection (DPMSD) model introduced in Chapter 2. In this section we shall assume that the parameters of this model correspond to a useful degree with familiarity and recollection. We make a number of arguments in favour of interpreting the parameters in this way, albeit cautiously, which are discussed in more detail in Chapters 2 and 4. For now, however, it is important to bear in mind that the data also fit well to a simpler unequal variance signal-detection (UVSD) model, the interpretation of which is more flexible and complex. Should this more parsimonious model accurately reflect the underlying processes, rather than simply approximate the resulting data efficiently, the following conclusions about familiarity and recollection in each condition would not be valid.

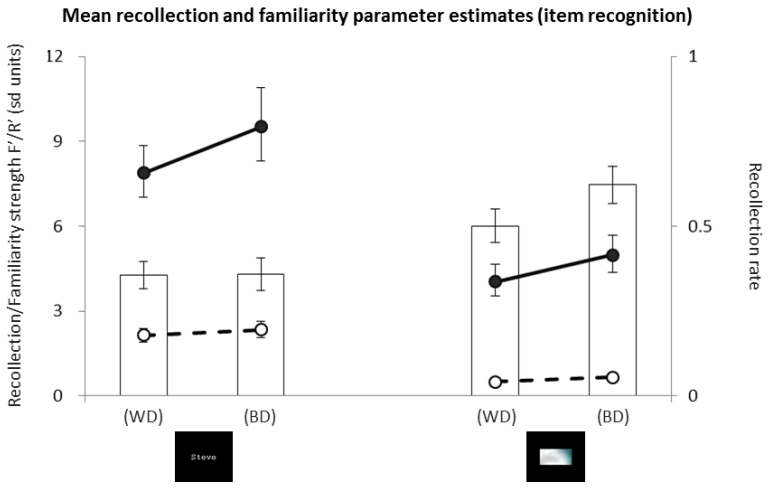
According to hierarchical likelihood ratio tests, allowing a familiarity signal to contribute to item recognition significantly improved the fit of the model [$\chi^2(144) = 1147.7; p < .001$], but including a similar familiarity component for the associative recognition conditions did not [$\chi^2(108) = 58.5; p > .999$]. Allowing the variance of confidence on recollected trials to differ from that on non-recollected trials further improved the fit [$\chi^2(36) = 149.9; p < .001$]; recollected trials were associated with a smaller range of confidence responses than were guessed or familiar trials. There was some evidence that this value might differ depending on whether the task was item or associative [$\chi^2(36) = 64.1; p = .003$], but paired t-tests on the estimates[1] of $v(R)$ when they were allowed to differ in this way showed no significant difference [$associative = 0.73, item = 0.72, t(35) = 0.52; p = .608$]. It is therefore difficult to say whether this represents a real (but inconsistent) difference or simply over fitting.

Mean estimates of each crucial memory parameter for the associative recognition

[1]The parameter $v(R)$ is a ratio and always greater than zero; its values follow a lognormal, rather than normal, distribution across participants. We therefore report the geometric rather than arithmetic mean, and perform statistics on the log-transformed values, which are approximately normally distributed. Values of recollection strength, d'_R, are treated in a similar way for the same reason. In contrast, some values of item familiarity strength, d'_F, were not greater than zero, therefore we report the arithmetic means for familiarity strength and perform statistics on the data without log-transforming. Similar results are obtained when statistics are performed on the log-transformed values after participants with $d'_F < 0$ in the condition of interest are excluded.



(a) Mean parameter estimates for the associative DPMSD model.



(b) Mean parameter estimates for the item DPMSD model.

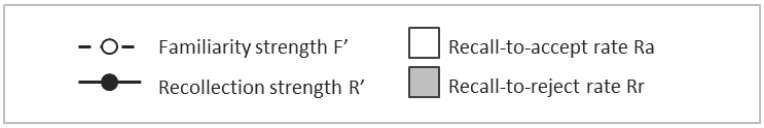


Figure 6.9: Familiarity and recollection estimates from the DPMSD model. (a) Familiarity does not reliably contribute to associative discrimination and recollection is necessary for above-chance performance. Better discrimination of name-name or between-domain than image-image pairs is driven primarily by more frequent, but not stronger, recollection. (b) In contrast, discrimination of individual items is supported by both recollection and familiarity. Both processes are stronger (provide more diagnostic information) for names than for images, yet recollection occurs more frequently for images, especially when they are studied as part of a between-domain pair.

task are presented in Figure 6.9(a). Repeated measures ANOVA with a single factor of condition (*WD-names*, *WD-images*, *between-domain*) revealed no significant difference in recollection strength across the three associative recognition conditions [$F(1.47, 51.54) = 2.61; p = .098$]. In contrast, recollection rate did vary across conditions. Two-way repeated measures ANOVA with factors of condition (*WD-names*, *WD-images*, *between-domain*) and recollection type (*recall-to-accept*, *recall-to-reject*) found significant main effects for both, reflecting less frequent recall-to-reject than recall-to-accept [$F(1, 35) = 52.13; p < .001$] and less frequent recollection overall for image-image pairs [$F(1.79, 62.47) = 17.42; p < .001$]. A marginal interaction [$F(1.98, 69.40) = 2.94; p = .060$] suggested recall-to-reject may have been more severely impaired for image-image pairs than recall-to-accept.Compared to both name-name and between-domain pairs, image-image pairs were less frequently recollected, but intact/rearranged discrimination was no weaker when recollection did occur.

Mean estimates of crucial memory parameters for the item recognition task are shown in Figure 6.9(b). A paired t-test on recollection strength d'_R for names suggested that it did not differ as a result of encoding condition [$WD = 7.88, BD = 9.52, t(35) = 1.63; p = .112$]. Similarly, encoding condition did not significantly influence the strength of recollection for individual images [$WD = 4.05, BD = 4.98, t(35) = 1.76; p = .087$], familiarity for names [$WD = 2.14, BD = 2.35, t(35) = 1.13; p = .266$] or familiarity for images [$WD = 0.50, BD = 0.66, t(35) = 1.12; p = .271$]. Fixing these four parameters across encoding condition did not significantly impair the fit of the model [$\chi^2(144) = 127.8; p = .830$]. The recollection rate R_a for names did not vary with encoding condition [$WD = 0.36, BD = 0.36, t(35) = 0.05; p = .961$], consistent with the absence of an overall performance difference across these conditions. By way of contrast, images were more frequently recollected if they had been encoded as part of a between-domain pair [$WD = 0.50, BD = 0.62, t(35) = 2.34; p = .025$]. In summary, the strength of familiarity and recollection were dependent only on item type; both were stronger for names than for images. In contrast, however, the rate of recollection was also dependent on relationship type; images were more frequently (but no more strongly) recollected when they were studied as part of a between-domain pair. According to the model, it is this increase in recollection rate which provides the basis for the result in 6.3.2: better discrimination of images studied as part of

between-domain pair.

6.4 Discussion

Here we replicate a key finding from Chapter 5. A regression analysis of associative discrimination performance revealed that it was predicted not only by item type, but also relationship type: between-domain pairs were more easily discriminated than within-domain pairs, even after counterbalancing component items across conditions. In this experiment, we also find that individual images, but not names, were better remembered when they were studied as part of a between-domain pair, raising the possibility that recognition for the same components may in fact differ across encoding conditions. Critically, however, careful analysis of the data however strongly suggests that this effect is not sufficient to explain the observed performance advantage for between-domain pairs.

6.4.1 Recognition of within and between-domain pair components

Component recognition was greater for images studied as part of a between-domain pair than those studied as part of a within-domain pair. In other words, a component-level effect exists that is caused by relationship type, and is not determined only by the type of item. In the introduction, we predicted that reducing the number of items of a particular type in each study list (the material-specific list length) might improve recognition of items of that type. If this effect was present in our experiment, intermixing conditions in each block should halve the material-specific list length for items encoded as part of a within-domain pair, improving their recognition as a result, but have no effect on memory for components of between-domain pairs. Analysis revealed that this was not the case; component recognition was unaffected by the total number of similar items in each study list. Even after intermixing conditions the relative recognition advantage for between-domain images remained just as strong.

Interestingly, Criss and Shiffrin (2004) found that for associative recognition, the number of similar pairs (rather than the number of similar components) was the

biggest determinant of performance. Perhaps in that study the pairs were encoded holistically, so that each pair was encoded, stored and retrieved as a single inseparable unit. Alternatively, the link or association between two components might have been encoded as a distinct piece of information, in addition to the components themselves. In this case the number of similar pairs, and therefore associations, might have the largest effect on performance if retrieval of components was considerably easier than retrieval of the associations. Just as there is no evidence of a material-specific list length effect for components in our study however, we also find no evidence of a list-length effect for pairs or associations. Intermixing conditions within blocks reduced the study list length of each pair type from 30 to 10, but this did not improve performance in any condition.

Why then do images show better recognition after being encoded as part of a between-domain pair? One possibility which is more consistent with the data is that the combination of stimulus types may allow more efficient encoding than when similar items are co-presented. Specifically, beyond a certain point the allocation of extra study time for images might be more valuable than for names, thereby leading to greater improvement in overall recognition. If this were the case, when the two different items were presented together attention and other resources could be allocated unevenly, in such a way as to maximise overall recognition. This could be tested directly in future, for example using eye-tracking, to determine how participants allocate resources at study and the effect this has on memory performance.

6.4.2 Material-specific differences in item recognition

Recognition for names and images were differently affected by the encoding condition. We have suggested one possible reason for this, namely differential allocation of attention or other resources critical for encoding when studying name-image pairs. This might reflect a difference in encoding rates across conditions, with names requiring less study time than images to reach an equivalent level of performance. Does recognition for names differ from recognition for images in other ways however? A principal components analysis of participant discrimination scores across 7 different tasks clearly separated item recognition for names and images (Figure 6.7). Variance in recognition of individual names was largely

(>70%) explained by the first component of the analysis, regardless of whether the name was studied as part of a within or between-domain pair. In contrast, variance in old/new discrimination for images was unrelated to the first component and instead was primarily (>65%) accounted for a second component. A similar pattern was observed after both VARIMAX rotation (which maximises uniqueness of variables for each factor) and QUARTIMAX rotation (which maximises uniqueness of factors for each variable). These results overall suggest a greater correspondence between performance when conditions share stimuli than when they share a task; associative discriminations for name-name and image-image pairs are more closely related to recognition of their components than to each other.

This pattern is not necessarily surprising, but it does suggest that the cognitive resources supporting old/new discrimination might differ depending on the type of stimulus being recognised. Material-dependent effects have been found in a wide range of previous studies (e.g. Galli and Otten, 2011, MacKenzie and Donaldson, 2007, Yick and Wilding, 2008). This is an important point to bear in mind given that a large proportion of memory studies use lexical stimuli exclusively.

In contrast, the components contributing to item recognition did not vary with study condition. In other words, although images were recognised more easily when they were studied as part of a between-domain pair, the principal component analysis does not provide any evidence that this reflects a qualitative difference in how they were encoded or retrieved. This is consistent with the theory that individual images were studied for longer on average when they were part of a between-domain pair. This would lead to better encoding and therefore higher recognition performance, without employing different encoding strategies or retrieval processes.

6.4.3 Recognition of names, but not images, predicts associative recognition performance

The largest two factors extracted by PCA appeared to be related to the presence of a name or an image respectively. More interestingly, the two factors are related in different ways to associative recognition memory. The first factor predominantly explained name recognition but was relatively closely related to

associative recognition performance. By contrast, the second factor was clearly image-related, yet it explained less of the variance in performance even for pairs composed entirely of images. In fact, while the majority of variance in item and name-name associative discrimination is explained by just two large factors, variance in between-domain and image-image associative recognition is incompletely explained, suggesting that other significant effects on memory may exist specifically for these two conditions.

A similar story emerges when performance across tasks is compared - associative recognition performance correlates significantly with name but not image recognition, regardless of whether the pair being associated consists of images, names or a mixture of the two. Component recognition seems to be important, but nevertheless provides a fundamentally incomplete explanation of associative recognition performance.

One way to interpret this pattern is that associative recognition in this study relies directly on name recognition. This is clearly likely to be true for name-name or between-domain pairs, which contain at least one name, but it is far less obvious that recognition of names, rather than images, should directly affect associative discrimination of image-image pairs.

Alternatively, associative recognition might be indirectly related to name recognition by relying on an overlapping set of cognitive resources. By this view, the largest extracted factor corresponds not to the presence of a name or type of task, but the efficacy of an underlying memory process which contributes both to item and associative recognition. Associative and item recognition are not clearly separated by either factor analysis or a distance map based on performances. In particular discrimination for name-name pairs is strongly correlated with recognition of individual names, consistent with the view that overlapping memory processes are used for recognition of both items and associations.

6.4.4 Dissociating recollection rate, recollection strength, and cued recall

If some subset of memory processes are shared between name and associative recognition, but less so with image recognition, what might these be? According

to the estimates from the dual-process mixture signal detection model, the same pattern appears in this chapter as the previous one: recollection was critical for associative recognition whereas familiarity did not contribute to discrimination to any significant degree. From a dual-process perspective, it might therefore be tempting to make the prediction that recollection was engaged in both associative and name recognition, while recognition of images was more heavily dependent on familiarity. In actual fact, according to the DPMSD model estimates the picture is more complex; the rate of recollection was higher for individual images than names, while the strength of familiarity was lower.

To better understand the relative contribution of each process in this experiment it becomes important to dissociate the rate of recollection from its diagnostic strength, something the DPMSD (but not DPSD or UVSD) model allows us to do. The two properties show distinct patterns when recollection and familiarity estimates are drawn in Section 6.3.6. Firstly, although recollection occurred more frequently for individual images than for names, the diagnostic strength of successful recollection was in the opposite direction; it was lower for images than for names. Furthermore, we demonstrated that recollection strength was consistent within item type but across encoding conditions. In other words, recollection provided more diagnostic evidence for names than for images, but this strength was the same regardless of whether the item was studied as part of a within or between-domain pair. In contrast, the rate of recollection - at least for images - did change with encoding condition, and it was this change which appeared to drive the improvement in component discrimination for between-domain pairs. Finally, all three associative recognition conditions were associated with similar recollection strength, with overall discrimination differences between conditions again being primarily driven by differences in recollection rate.

Do these dissociations make sense in terms of dual-process theory? To answer this question it may make sense to consider recollection strength as a property that is closely related to the type of stimulus being recollected. When recollection occurs, some details about the original study episode are made available, but as both we and others have shown (Chapter 4; citealtMickes2009,Slotnick2010,Hautus2008), the accuracy of these details may vary considerably. Individual names are associated with some pre-experimental familiarity, while the images are arguably more complex and certainly more novel. Recollection of an image might, as a result, be

less precise or more prone to error than recollection of a name. For example, one may recollect that a light blue image was presented at study without being able to bring it more precisely to mind, this might therefore provide weak diagnostic evidence for a specific light blue image presented at test being old rather than new. Accordingly, although familiarity is weaker for images, it may be relatively more important when making an old/new decision, because recollection rarely provides very strong discrimination on its own.

Conversely, a name such as 'Brian' might be recollected accurately most of the time and only occasionally confused with a similar name, overall leading to stronger diagnostic evidence both for item and associative recognition, and less reliance on familiarity. This may arise as a result of of the pre-experimental familiarity of the stimulus. According to some neural network models of memory (Norman and O'Reilly, 2003) repeated presentation of a stimulus results in a more distinct, pattern separated representation being formed in sparse networks. Since the representation overlaps less with other stimuli it is confused with them less frequently, leading to more accurate (stronger) recollection.

A participant's ability to recollect details accurately may thereby correlate with both name and associative recognition performance, but less so with image recognition where more of the variance in performance is due to differences in familiarity across participants. This account is also consistent with the fact that successful recollection, in particular recall-to-reject which should require relatively strong and accurate recollection, was more frequent for pairs containing at least one name than image-image pairs. Interestingly, by this view cued recall rates may vary across stimulus categories in a similar way (and for similar reasons) as recollection strength, but in the opposite direction to recollection rate. Recall-to-reject rates specifically rely on (accurate) cued recall of a component not shown at test, and were dramatically lower for image-image pairs than for pairs containing at least one name, driving much of the difference in overall discrimination between these conditions. This suggests that images may have been harder to recall from memory accurately enough to identify them as being originally paired with one of the components presented at test. Given their pre-experimental novelty, abstract nature and complexity this seems reasonable, but note that it is in the opposite direction to recollection (recall-to-accept) rates for individual images presented at test. The difference arises because the richness of a component, which provides

a large number of cues for recollection when it is shown at test, becomes a disadvantage when it is not presented, and is required to be recalled from memory. We shall return to this issue in more detail in the general discussion of this thesis.

The results arguably therefore fit with dual-process theory, with the caveat that model choice is fundamental to any conclusions about recollection and familiarity. For example, they appear to support the widely held view that associative and item recognition likely rely on some similar, overlapping resources: in particular recollection may support both item and associative recognition (Yonelinas, 2002a). They also fit with the view that other processes, such as familiarity (or perceptual fluency; Yovel and Paller, 2004) are specifically diagnostic in old/new item recognition, and provide little advantage when the task involves the recognition of associations (Donaldson and Rugg, 1998). Importantly, they suggest that the strength of recollection might be dissociated from the frequency with which it occurs. Interestingly, the former may, at least in this study, be predominantly determined by the properties of the stimulus being recollected, while the latter might be more sensitive to experimental factors such as encoding time. Similarly, the results also highlight the importance of distinguishing recall from recollection, and suggest that the two may dissociate as a function of stimulus category. It is worth noting, however, that while this dissociation explains the relatively low recall-to-reject rates for image-image pairs, it also predicts reduced recall-to-reject rates for between-domain compared to name-name pairs. In actual fact, as is clear from Figure 6.3, recall-to-reject rates for between-domain pairs are no lower than for name-name pairs. Once again component properties do not reliably predict performance for between-domain pairs.

6.4.5 Component differences do not explain the recognition advantage for between-domain pairs

Process estimates are not the only source of evidence that between-domain pairs are associated with an unexplained discrimination advantage. Correlational and factor analyses together suggest that associative discrimination of name-name pairs is strongly related to component recognition, whereas associative discrimination of between-domain pairs is not, despite similar overall performance. If less variance in between-domain associative recognition can be explained by recog-

nition of the individual components, this in turn suggests that performance is determined by additional factors besides the properties of these components. Another important finding is that discrimination for between-domain pairs is at best only weakly predicted by recognition of individual images. Given that only images, and not names, are better recognised when encoded as part of a between-domain pair, this provides a further argument that component recognition differences cannot sufficiently account for the phenomenon of improved between-domain pair recognition.

Indeed, when differences in component recognition were accounted for by using old/new component discrimination instead of item type to predict associative discrimination, a significant effect of relationship remained. The components of between-domain pairs may be easier to recognise than those of within-domain pairs, but the pairing of a name and an image imparts an even greater - and unexplained - associative recognition advantage. What then, might this advantage stem from, and why does recognition of between-domain pairs relate so loosely to performance on other tasks? One possibility, investigated in the next chapter, is that between-domain pairs are encoded and retrieved in a fundamentally different way to the other pair types in this experiment.

Chapter 7

Unitization

The findings from Chapters 5–6 highlighted a reliable effect of component relationship on associative episodic recognition. Specifically, greater differences between component items (as found in between-domain pairs) promoted better discrimination of intact pairs from recombined lures. In the previous chapter we further demonstrated that this improvement in discrimination could not be explained purely by differences in recognition for individual components across conditions. An effect of relationship exists independently from component recognition effects. Together these finding provoke the question we pose in this chapter: By what mechanism might the relationship between items - independently of the items themselves - affect how they are encoded or retrieved from memory? Although the domain dichotomy account can be ruled out as an explanation, since it predicts the opposite effect (i.e., *impaired* memory for between-domain pairs), an alternative theory of unitization might plausibly explain this change in memory performance.

7.1 Introduction

Unitization refers to the perception and processing of multiple associated items as a single coherent whole, rather than as distinct, explicitly associated components. Unitization is currently a concept of considerable interest to memory researchers (Bader et al., 2010, Haskins et al., 2008, Jäger et al., 2006, Quamme et al., 2007, Rhodes and Donaldson, 2007). One reason for this is that it provides an expla-

nation for the apparent phenomenon of associative familiarity: the recognition of a particular combination of items, rather than only the items themselves, on the basis of familiarity. Until relatively recently, discrimination of intact from rearranged combinations of items (e.g., if A-B and C-D are studied pairs, then A-B is an intact pair but A-C is rearranged) was thought to rely overwhelmingly on explicit recollection of the original presentations and their associated links (Donaldson and Rugg, 1998, Hockley and Consoli, 1999, Yonelinas, 1997). More recently, however, several studies have suggested that if multiple items are to some extent unitized at encoding, recognition of their associations may be based at least partly on familiarity (Quamme et al., 2007, Rhodes and Donaldson, 2007). By this view, under certain circumstances unitization influences the qualitative processes used to support episodic retrieval.

7.1.1 Perceptual unitization

According to the definition above, unitization can occur for various stimulus types, by different encoding strategies and at different levels of representation or stages of processing. For the purposes of investigating and understanding unitization, it might therefore be useful to characterise unitization more rigorously, or divide it into more strictly defined subtypes. Most straightforwardly, given that it is a process which is purported to influence episodic memory, we can define episodic unitization as being able to occur after a single study presentation.

More precisely, we might wish to separately identify and investigate subtypes of unitization based on the level at which they occur. For example, two separate words such as 'sea' and 'cube' might be conceptually or semantically unitized by defining a new compound word 'seacube: a cube of salt water', which can be encoded and recognised as a single item. Unitization for words has been demonstrated in a number of studies (Bader et al., 2010, Haskins et al., 2008, Quamme et al., 2007, Rhodes and Donaldson, 2007; 2008). Two images, in contrast, might be perceptually unitized by perceiving and encoding their combined presentation as a single more complex image, again leading to recognition of the combined image independent of any recollection of its components. This perceptual form of unitization has not been as widely demonstrated as conceptual unitization of words (though see Yonelinas et al., 1999 and Diana et al., 2007), but it could

potentially provide a more general basis for unitizing associated components, particularly in terms of the different classes of stimuli it could operate on. Alternatively, perceptual unitization might rely heavily on the component items fitting together into a familiar structure, such as the features of a face (Yonelinas et al., 1999). One interesting unanswered question therefore is the level of control a viewer has over episodic unitization: to what extent is the perception of a 'unit' dependent on the appearance of the to-be-unitized stimuli and the prior experience of the viewer? The question of whether unitization is restricted to specific stimuli or can be employed more generally should be addressed in future studies.

7.1.2 Is unitization a plausible explanation for improved between-domain recognition?

Having defined what we mean by episodic unitization, we now consider whether it might provide a plausible explanation for improved recognition of between-domain pairs. Firstly, the recognition of associated pairs should be aided by unitization, since it purportedly allows recognition based on both familiarity and recollection. If the familiarity supported by unitization is relatively weak, some might argue that it would provide little additional diagnostic ability, being instead frequently ignored in favour of stronger recollection (see for example Yonelinas, 1994, though Wixted, 2007a favours an alternative view that signals from multiple processes are generally combined). Nevertheless, in this case familiarity should still be useful on trials for which recollection fails, or recollection alone may provide a stronger basis for discrimination of unitized pairs. For example, a unitized pair of items might be represented by more overlapping or closely linked representations than a non-unitized pair, promoting pattern-completion and retrieval of both elements of the episode. This is in essence the domain dichotomy argument, except that it is not restricted to perirhinal cortex, and representational overlap is dependent on encoding strategy rather than stimulus similarity.

Regardless of hypothesised mechanisms, results in the unitization literature give good reason to expect that unitization should be accompanied by better recognition memory. Recent published reports claiming to have manipulated unitization in healthy participants almost invariably show improved memory, and particularly hit rates, for the unitized relative to non-unitized conditions (e.g. Bader

et al., 2010, Diana et al., 2008, Haskins et al., 2008, Jäger et al., 2006, Kounios et al., 2003, Opitz and Cornell, 2006, Rhodes and Donaldson, 2007; 2008; though see also Ford et al., 2010 for an exception).

The improved associative discrimination of between-domain pairs observed in the preceding chapters might therefore be explained by more robust unitization of their components at study, compared to name-name or image-image pairs. This would give rise to the observed pattern of improved associative recognition, independently from recognition of their components. Thus, unitization seems to provide a plausible hypothesis for the pattern, but in order to test this hypothesis we require a means of measuring the extent of unitization in each condition.

7.1.3 Detecting unitization

When can we be sure that unitization has occurred? From a subjective, phenomenological perspective one might argue on the basis of plausibility: two or more items might 'feel' like they form a new, holistic representation. More objectively, it has also been argued that unitization of multiple items is necessary for that association to be recognisable on the basis of familiarity (Diana et al., 2007). By this view, evidence of familiarity for a group of items, at least where those items might plausibly have been unitized, would constitute evidence for the process of unitization itself.

Notwithstanding the difficulty involved in determining whether familiarity is present for a given task, it is difficult to avoid circularity if familiarity provides the evidence for unitization, given that unitization has itself been invoked as a post-hoc explanation of evidence for familiarity in the form of curvilinear ROCs (Yonelinas, 1997). Some attempts have been made to define a more objective test: Mayes and colleagues (2007) suggested that unitization could be defined by demonstrating measurable costs as well as benefits to memory. More precisely, unitization could be said to have occurred to the extent that memory or perception is strengthened for the associated pair, but correspondingly weakened for the individual components.

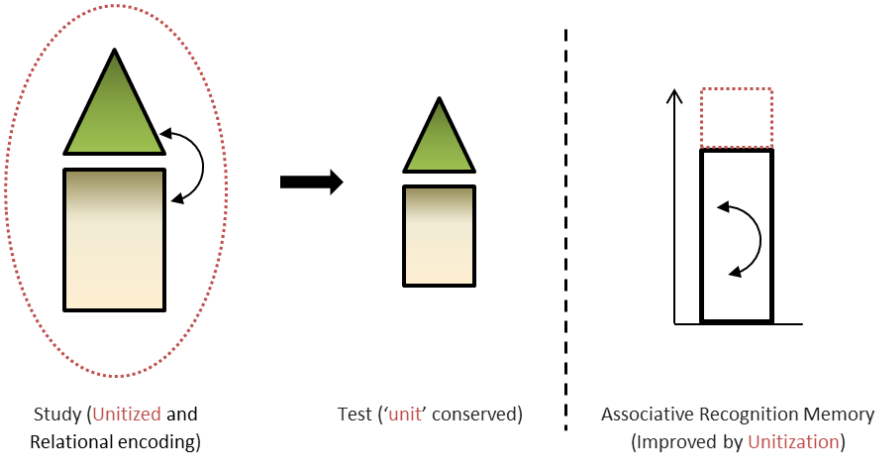
In practice, it may prove difficult to isolate a specific unitization-driven deficit for component recognition. For example, a deficit for component familiarity caused by unitization could plausibly be counterbalanced by improved recollection of

the component items, achieved by using the familiar, unitized representation as a cue. Clearly any attempt to accurately predicts and measure costs is itself fraught with difficulty and to our knowledge no published reports have yet demonstrated reliable costs. Thus, we focus instead on an alternative predicted property of a unitized pair: an increased reliance upon study-test consistency.

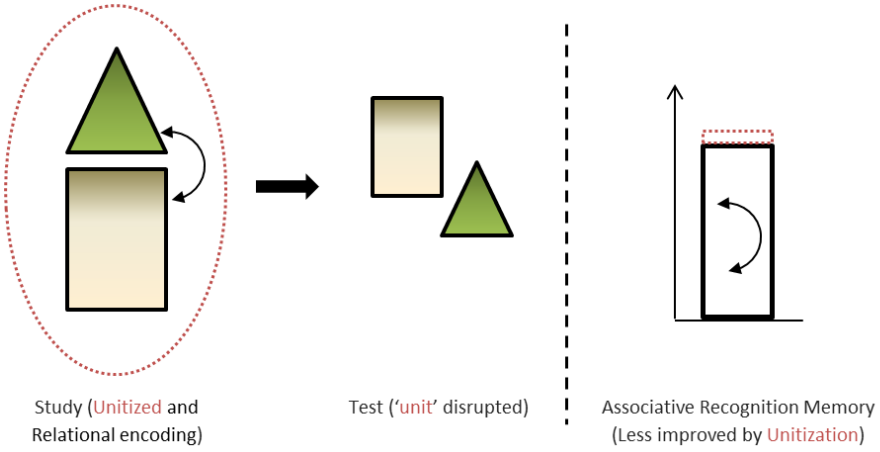
7.1.4 The perceptual-switch

By its most basic definition, unitization improves the recognition of two or more associated items by treating their presentation as a single 'unit' (Figure 7.1(a)). It follows that disrupting the integrity of this unit at test, while keeping the abstract association between its components intact, should impair recognition based upon unitization more than recognition based upon explicit recollection of the items and their associations (Figure 7.1(b)). A similar logic has been applied in an attempt to measure unitization before (Haskins et al., 2008, Speer and Curran, 2007). Haskins et al. (2008) demonstrated that reversing the order of components of a compound word (e.g. 'birdbath' becomes 'bathbird') impaired recognition of the associated words in conditions where a conceptual form of unitization was hypothesised to operate. Using a related but subtly different approach, Speer and Curran (2007) attempted to encourage or disrupt the perceptual unitization of pairs of fractals by either fixing or varying their relative positions across multiple study presentations. One disadvantage resulting from the use of multiple study episodes, however, is that recognition becomes less clearly episodic. Unitization could still occur for individual study presentations, leading to differences mainly in the number of stored representations rather than the way in which they are represented. Furthermore, recollection rate and strength are likely to be affected by both the number of relevant study episodes and in particular their heterogeneity, leading to differences across conditions which may be hard to separate from purported effects of unitization.

In this experiment therefore we present components no more than once at study, employing a similar approach to that used by Haskins et al. (2008) (see the supplementary online data for that article). That experiment disrupted the study-test overlap in order to detect largely conceptual or semantic unitization, however the between-domain pairs which demonstrated improved recognition in our study



(a) Intact study-test overlap allows perceptual unitization to aid recognition of novel associations.



(b) Disrupted study-test overlap makes recognition more dependent on explicit recall of associations.

Figure 7.1: An illustration of the potential role - and selective disruption - of unitization in associative recognition. The association between two component items can be encoded both as an explicit relationship between two items (black curved arrow) and a holistic or 'unitized' representation of the presentation as a single object (red broken circle). (a) If at test the components are presented in the same arrangement, memory may be based on recognition of the unitized whole as well as recollection of the relationship, and therefore improved. (b) If instead the components are presented in a different arrangement, unitization may help less since the 'unitized' representation does not reappear at test, but memory based on explicit recognition of the individual components and their relationship should be less strongly affected.

comprised abstract images as well as names, making a conceptual unitization explanation unlikely. Instead, therefore, we consider the possibility that these pairs were more robustly unitized at a perceptual level. If this is the case, a switch disrupting the perceptual appearance should impair performance on these pairs, and to a greater degree than for within-domain pairs (since within-domain pairs are predicted to be less robustly unitized). The perceptual-switch manipulation is illustrated in Figure 7.2. Here the spatial position of component items within a pair is reversed between study and test, so that items presented above central fixation at study appear below it at test, and vice-versa. Since the individual components do not change their orientation, and move independently, they are less likely to be perceived, encoded and recognised as the same single object at study and test - by definition they are less unitized.

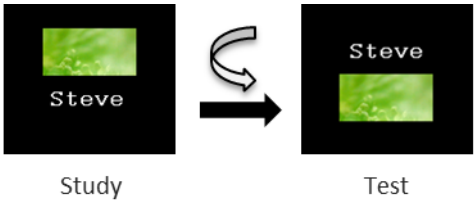


Figure 7.2: The perceptual-switch. In this experiment the perceptual-switch was applied to half of all pairs at test; for these the above-fixation component at study was presented below-fixation and vice-versa. The perceptual-switch was applied equally frequently to intact and rearranged test pairs.

7.1.5 Does unitization explain better recognition for between-domain pairs?

Here we test the extent to which associative recognition of within- and between-domain pairs might be supported by perceptual unitization. Between-domain pairs have been shown to be recognised more readily than within-domain pairs (Harlow et al., 2010). Does unitization play a role in recognition of either pair type, and do differences in unitization provide an explanation of differences in recognition?

We compared associative recognition performance for pairs which were kept perceptually constant between study and test, to performance for pairs which were

perceptually-switched. We found that recognition was impaired by the perceptual-switch, and that this effect was specific to between-domain pairs, consistent with the hypothesis that better recognition for between-domain pairs was a least partly driven by greater unitization than for within-domain pairs. We then investigated whether the perceptual-switch, and by extension unitization, influenced memory by operating on familiarity or on recollection. Using a DPSD model, as in previous unitization studies, suggested that the perceptual-switch selectively impaired associative familiarity, consistent with previous descriptions of unitization. Importantly, however, when a more accurate DPMSD model was used to analyse the data, no evidence of familiarity was found in any condition, and the perceptual-switch was found to selectively impair the probability of recollecting an intact pair at test. The results therefore provide a possible alternative characterisation of unitization as improving recollection even when familiarity is unaffected, and also clearly demonstrate that the choice of an appropriate model to analyse confidence data has a significant effect on the conclusions drawn.

7.2 Methods

Participants studied pairs of items and later discriminated between intact and rearranged pairs of the same items, rating their confidence from 1-5 after each decision. Half of all pairs (both intact and rearranged) were perceptually-switched across study and test, Figure 7.2. Memory performance was examined using 9-point ROC curves, constructed separately for each participant. The contribution of familiarity and recollection to performance was estimated from both the commonly-used DPSD model (Yonelinas, 1994) and the graded DPMSD model suggested by the results of Chapter 4.

7.2.1 Participants and procedure

Eighteen right-handed participants (8 female; mean age 20.8, range 18-27) completed the experiment and all data sets were included in the main analyses. The procedure was similar to that used in Chapter 5, Experiment 1 and is summarized in Figure 5.1. The experiment was divided into 12 randomly-ordered blocks, 4 for each stimulus condition, and each block was further divided into a 36-trial

study phase and a 24-trial test phase. At test, 12 pairs of items were intact (appeared together in the preceding study phase) and 12 were rearranged (appeared in separate study trials). Half of the test pairs (6 intact, 6 rearranged) were perceptually-switched between study and test. Thus, in total 24 test pairs in each of the three stimulus conditions were intact and perceptually-switched, 24 were rearranged and switched, 24 were intact and not switched, and 24 were rearranged and not switched. All four types of test pair were randomly intermixed during each test block, and study pairs were presented in random order.

At both study and test the mapping of left and right buttons to (*intact*, *rearranged*) and (1-5) responses was fully counterbalanced across blocks of 4 participants; the stimulus condition (*WD*, *WD*, *BD*) and test condition (*intact*, *rearranged*, *not shown*) of each item was fully counterbalanced across blocks of 9. On average the procedure took 2 hours to complete, including a practice block and debriefing.

7.2.2 Stimuli

We employed the three stimulus conditions used in the previous two chapters, Figure 5.1. Uniquely for this experiment however, half of all presentations (for all three conditions, and for both subsequently intact and subsequently rearranged pairs) were perceptually-switched, Figure 7.2. For these pairs, components displayed above fixation at study were displayed below fixation at test, and vice-versa. It was explained clearly to participants before commencing the experiment that the half of the pairs would be switched in this way, but that their task was simply to distinguish intact from rearranged pairs regardless of whether the component items had switched position. The pool of names and images used to construct stimulus pairs was identical to that used in the previous two chapters, and is described in Section 3.1.2.

7.2.3 Data analysis

We estimated the contribution of familiarity and recollection by fitting the confidence data to two different memory models, 1) the widely used dual-process signal detection (DPSD) model, (Yonelinas, 1994), and 2) the mixture signal-detection

(DPMSD) model outlined in Section 2.3.3. Both models provide estimates of familiarity strength, d'_F, as well as recollection rates separately for intact (recall-to-accept, R_a) and rearranged (recall-to-reject, R_r) pairs. The DPMSD, but not the DPSD model, also allows for recollection to vary in strength; it therefore includes an extra parameter per condition to account for this, d'_R, and an additional parameter per participant, $v(R)$, which defines the variance of this recollection distribution.

7.3 Results

We assessed associative recognition using the discrimination statistic d_a, calculated directly from participant confidence judgments. Overall performance is summarised in Figure 7.3. Visual inspection of the data suggests that discrimination was elevated for between-domain compared to within-domain pairs, name-name pairs were better recognised than image-image pairs, and that perceptually-switching pairs between study and test impaired recognition.

7.3.1 Factors affecting performance

We first tested whether the factors of item type (*0 names*; *1 name*; *2 names*), relationship type (*within-domain*; *between-domain*) and perceptual-switch (*fixed*; *switched*) independently affected performance on the task by performing a linear regression analysis. The analysis revealed a significant contribution for each factor of interest, confirming that item type, relationship type and the perceptual switch all had a measurable effect on performance. Item type had a significant effect on intact/rearranged discrimination d_a ($B = 0.725; p < .001$), reflecting improved discrimination for pairs comprising names rather than images. As also found in the previous two chapters, relationship type affected performance independently of item type, with better discrimination for between- than within-domain pairs ($B = 0.677; p < .001$). Finally, the perceptual-switch manipulation had the intended influence on memory; participants were less able to distinguish intact from rearranged pairs when the items had switched positions between study and test ($B = 0.246; p = .044$).

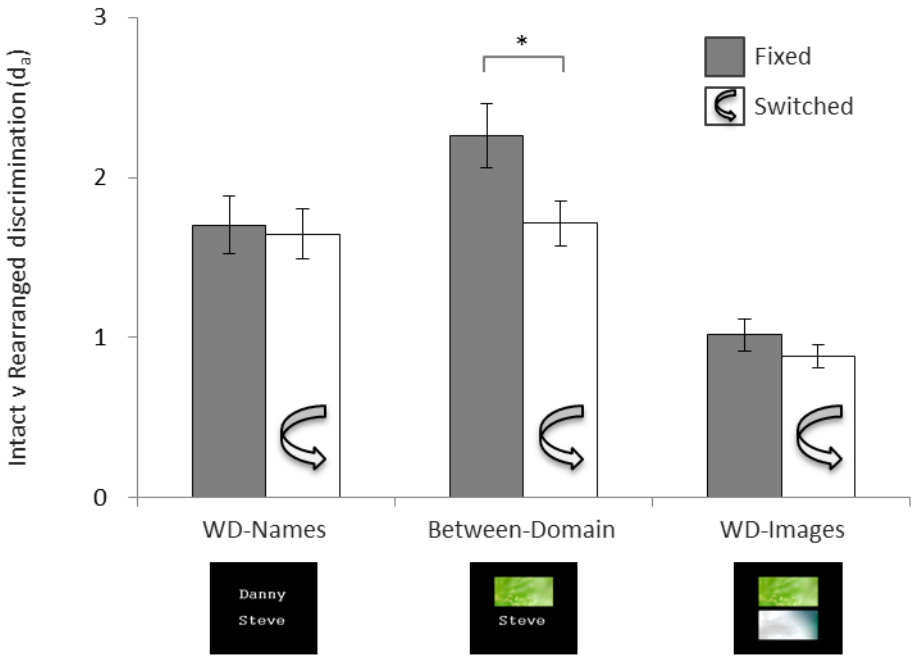


Figure 7.3: Discrimination of intact from rearranged pairs, d_a, as a function of stimulus condition and perceptual-switch. As found in previous experiments, discrimination was significantly affected by item type (names lead to better discrimination) but also relationship (between-domain lead to better discrimination). For half of all pairs the position of the component items was swapped between study and test, the perceptual-switch. The effect of this manipulation was to reduce discrimination, but only for between-domain pairs.

7.3.2 Condition-specific effects of the perceptual-switch on performance

Having established that the perceptual-switch impaired associative recognition, we then tested whether this effect was specific to particular item or relationship types. Using linear regression, we tested the effect of item type and relationship type as defined above on the efficacy of the perceptual-switch. That is to say, a new dependent variable was defined for each stimulus condition as the difference in d_a between perceptually-switched and non-switched pairs. Item type had no significant effect on the size of the perceptual-switch effect ($B = -0.079; p = .664$) reflecting the fact that the perceptual-switch did not disrupt pairs of names significantly more or less than pairs of images. However, the type of relationship did have a significant effect ($B = 0.453; p = .005$), with discrimination for between-

domain pairs more strongly impaired by the perceptual-switch. The constant term did not vary significantly from zero ($B = 0.135; p = .294$); this reflected the fact that the perceptual-switch did not have a general effect, but only a relationship-specific effect. To be clear, a constant term of zero in the analysis indicates that there was no evidence that the perceptual switch reliably affected performance for within-domain pairs.

Since the regression analyses used above are underpowered, a repeated-measures ANOVA was used to establish more directly how the perceptual-switch affected within- and between-domain pairs, taking into account the within-subject experimental design. For this purpose, name-name and image-image conditions were collapsed together for each participant by combining individual trials and recalculating discrimination d_a, forming a single within-domain condition comprising exactly the same component items (but twice as many trials) as the between-domain condition. Thus the two conditions could be directly compared to determine the effects of relationship type in the absence of confounding differences in component items. Discrimination for these two conditions is illustrated in Figure 7.4.

The data were submitted to a 2×2 repeated measures ANOVA with factors of relationship type (*within-domain*, *between-domain*) and perceptual-switch (*fixed*, *switched*). Consistent with the results of the regression analysis above, ANOVA confirmed significantly greater discrimination for between- than within-domain pairs [$F(1,17) = 59.21, p < .001$] and lower discrimination for perceptually-switched than fixed pairs [$F(1,17) = 15.13, p < .001$]. A significant interaction [$F(1,17) = 8.54, p = .009$] reflected a greater effect of perceptual switch on between- than within-domain pairs. Paired t-tests confirmed that the main effect of perceptual-switch on discrimination was driven entirely by a decrease in discrimination for between-domain pairs [$t(17) = 3.73, p = .002$], there was no reliable effect on within-domain pairs [$t(17) = 1.06, p = .305$]. Finally, a paired t-test on switched pairs only revealed that even after the perceptual-switch, between-domain pairs were still discriminated more easily than within-domain pairs of the same items [1.72 *vs* 1.18; $t(17) = 4.90, p < .001$].

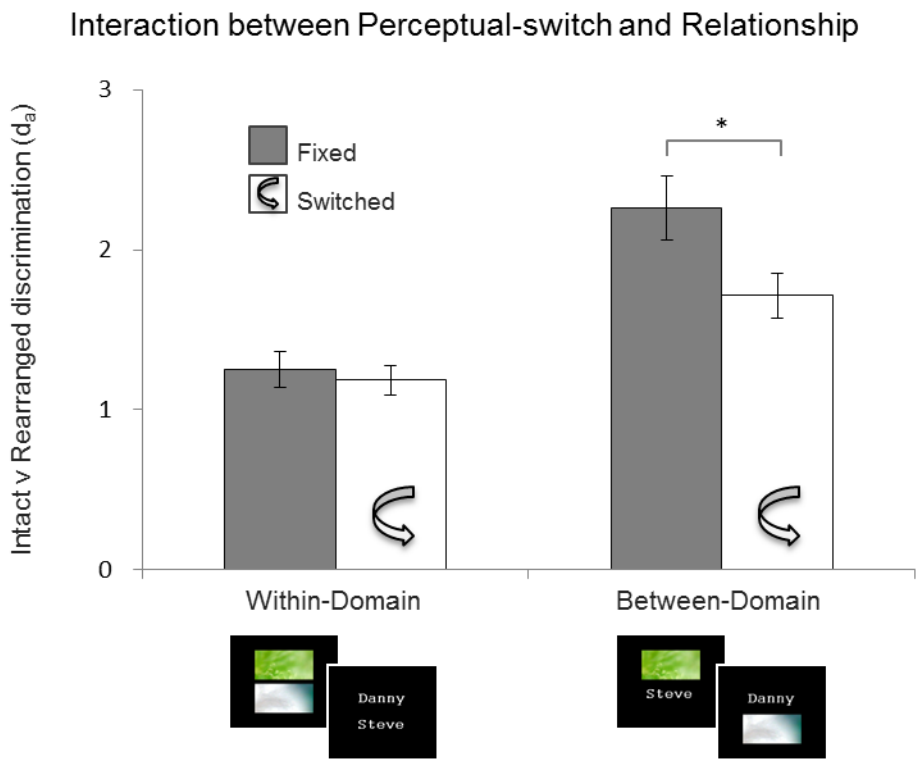


Figure 7.4: Interaction between perceptual-switch and relationship. Between-domain pairings of items were significantly less well recognised following the perceptual-switch, in contrast to within-domain pairings of the same items, which were virtually unaffected.

7.3.3 Familiarity using a DPSD model

We next tested whether familiarity contributed to associative recognition in each condition by fitting the confidence data to the associative DPSD model, which provides estimates of familiarity strength d'_F and recollection rates for, separately, intact $p(R_a)$ and rearranged $p(R_r)$ pairs (the probability of correctly recognising a pair as intact or rearranged by recalling the study presentation or presentations respectively).

Paired t-tests revealed what is apparent from inspection (Figure 7.5): d'_F was reliably greater than zero, meaning that according to the DPSD model familiarity contributed to associative recognition in each conditions, both before and after the perceptual-switch. This result was corroborated by a likelihood ratio test on the significance of the d'_F parameter, the inclusion of which significantly improved the fit over a recollection-only model $[\chi^2(108) = 240.66, p < .001]$.

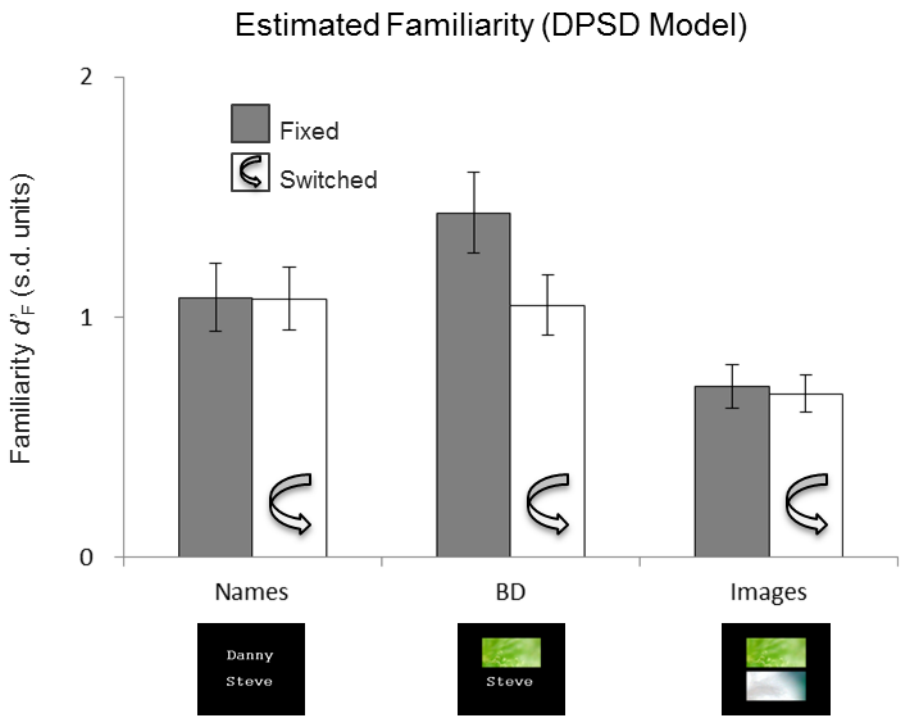


Figure 7.5: Estimated familiarity under a DPSD model. According to estimates from the DPSD model, the perceptual-switch had a strong effect on familiarity to between-domain pairs.

7.3.4 Familiarity using a DPMSD model

Under the assumption of the DPSD model that recollection is not significantly graded, familiarity appeared to contribute strongly to associative recognition. Was this also true when recollection strength was allowed to vary, as has been previously demonstrated is more likely (Mickes et al., 2009, Slotnick, 2010)? We fit confidence ratings from each participant to the DPMSD model, which provides estimates of familiarity strength d'_F, recollection strength d'_R, the variance of the recollection distributions $v(R)$, and recollection rates for, separately, intact $p(R_a)$ and rearranged $p(R_r)$ pairs.

First, the necessity of the extra parameters included in the DPMSD model was assessed directly. Including for each participant one variance parameter $v(R)$ and six (one per condition) recollection strength parameters d'_R significantly improved the likelihood of the observed data $[\chi^2(126) = 99.67, p < .001]$ compared to using the DPSD model. As predicted by previous studies, recollection was most accurately modelled as a graded process.

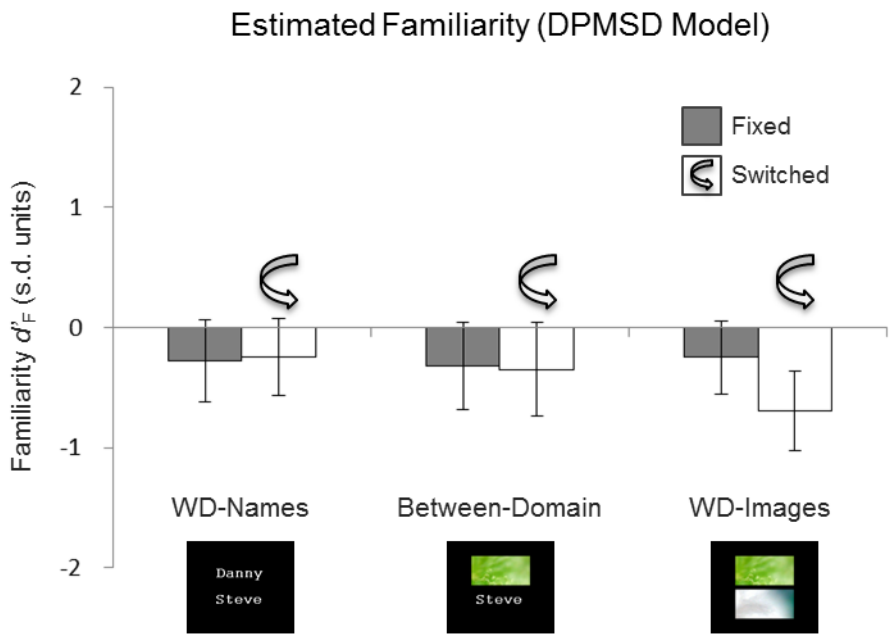


Figure 7.6: Estimated familiarity under a DPMSD model. In contrast to the conclusion drawn by the DPSD model, the (better-fitting) DPMSD model estimated no significant contribution from familiarity in any condition.

Second, the significance of the familiarity parameter d'_F was similarly assessed using a likelihood ratio test. In contrast to the DPSD model, the inclusion of d'_F gave no significant advantage over setting it to zero $[\chi^2(108) = 45.98, p = .865]$. By allowing d'_F to vary freely (above and below 0) we also recovered an estimate of familiarity contribution in each condition from the full, over-specified mixture model (Figure 7.6). There was no evidence that familiarity was contributing weakly, the trend was in the opposite direction to above-chance familiarity. Under this more accurate model, familiarity does not contribute to associative recognition in any condition.

7.3.5 Process-specific effects of the perceptual-switch using a DPSD model

We further investigated whether the perceptual-switch impaired discrimination by affecting familiarity or recollection in between-domain pairs. Within-domain pairs were not analysed in this way since the perceptual-switch was shown to reliably impair overall discrimination only for between-domain pairs, and so there

was no a priori reason to expect meaningful process-specific differences for within-domain pairs.

A paired t-test revealed that the perceptual-switch significantly reduced familiarity [$t(17) = 2.43, p = .027$] for between-domain pairs. Recollection rates were submitted to a 2×2 repeated measures ANOVA with factors of test condition (*intact*, *rearranged*) and perceptual-switch (*fixed*, *switched*). A main effect of perceptual-switch was not significant [$F(1,17) = 2.49, p = .133$], but a main effect of test condition [$F(1,17) = 5.89, p = .027$] reflected more frequent recollection for intact than rearranged pairs. This interacted with perceptual-switch [$F(1,17) = 5.89, p = .013$], such that the recollection advantage for intact over rearranged pairs was reduced when items switched positions between study and test. Thus, under the assumptions of a DPSD model the evidence that recollection is affected by the perceptual-switch is much weaker than the corresponding evidence for an effect on familiarity.

7.3.6 Process-specific effects of the perceptual-switch using a mixture model

The most parsimonious mixture model does not include familiarity, so under the assumptions of this model how does the perceptual-switch impair discrimination for between-domain pairs? A paired t-test revealed that the strength of recollection did not significantly decrease when pairs were switched [$d'_R(fixed) = 3.71, d'_R(switched) = 3.89; t(17) = 0.31, p = .763$]. All estimated values of d'_R were strictly positive and followed a lognormal, rather than normal, distribution. Comparing geometric means is more appropriate in this situation than comparing arithmetic means; however the geometric means of recollection strength were also not significantly affected by the perceptual-switch [$d'_R(fixed) = 3.37, d'_R(switched) = 3.52; t(17) = 0.33, p = .747$].

Recollection rates (Figure 7.7) were submitted to a 2×2 repeated measures ANOVA with factors of test condition (*intact*, *rearranged*) and perceptual-switch (*fixed*, *switched*). A main effect of perceptual-switch was significant [$F(1,17) = 5.20, p = .036$], but the frequency of recollection was not found to be more or less frequent for intact pairs [$F(1,17) = 0.75, p = .400$]. Nor did the perceptual-switch interact significantly with test condition [$F(1,17) = 2.68, p = .120$]. While the

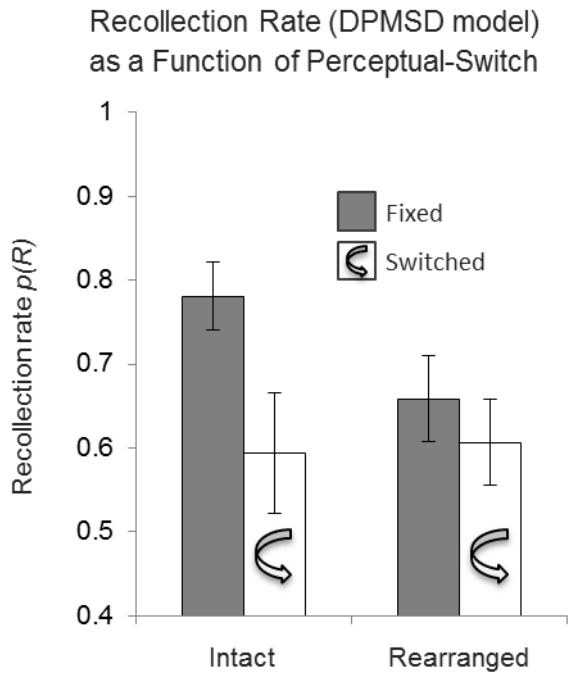


Figure 7.7: The effect of perceptual-switch on between-domain pair recollection. Perceptually-switching pairs, according to the DPMSD model, selectively reduced the rate (but not strength) of recollection.

DPSD model suggests the perceptual-switch impairs familiarity (with little effect of recollection), the DPMSD model instead ascribes reduced performance to a reduction in the rate of recollection (and no contribution of familiarity at all).

7.3.7 Effects of the perceptual-switch on confidence

Participants completed a relatively small number of trials per condition (24 intact & 24 rearranged), meaning that parameter estimates from the memory models analysed above should be treated cautiously (Yonelinas, 2002a). We therefore also analysed the effect of the switch on the confidence data directly (Figure 7.8). We used repeated measures ANOVA with factors of test condition (*intact*, *rearranged*), pair type (*WD-Name*, *WD-Image*, *BD*) and perceptual-switch (*fixed*, *switched*). This revealed main effects of perceptual switch [$F(1,17) = 22.11, p < .001$] and pair type [$F(1,17) = 77.39, p < .001$]; the perceptual-switch reduced confidence, and WD-Image pairs were less confidently recognised than WD-Name or BD pairs (both $p < .001$). Perceptual switch did not interact with pair type,

suggesting it had a comparable effect on confidence regardless of the type of stimuli, but it did interact strongly with test condition [$F(1,17) = 29.36, p < .001$]: while the perceptual-switch reduced confidence for intact pairs, it did not do so for rearranged pairs (and in fact the trend was for improved confidence to these pairs).

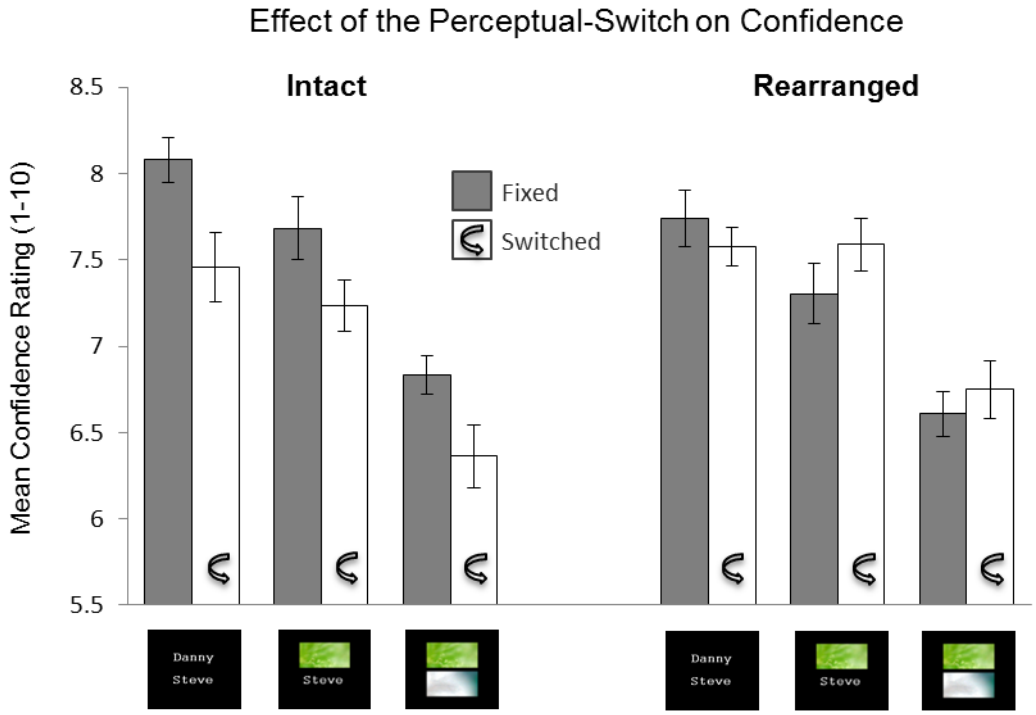


Figure 7.8: The perceptual-switch consistently reduced confidence for intact pairs, but had no significant effect on rearranged pairs.

7.3.8 Effects of the perceptual-switch on reaction times

We next analysed the effect of the perceptual-switch on reaction times for correct responses[1], summarised in Figure 7.9. We used repeated measures ANOVA with factors of test condition (*intact*, *rearranged*), pair type (*WD-Name*, *WD-Image*, *BD*) and perceptual-switch (*fixed*, *switched*). This revealed main effects of all three factors: intact pairs were recognised more quickly than rearranged pairs [$F(1,17) = 57.84, p < .001$], the perceptual-switch slowed response times [$F(1,17) = 10.77, p = .004$], and responses were slowest to WD-Name pairs

[1]The reaction time data was log-normally distributed; we therefore report geometric means and all statistical analyses are performed on the log-transformed data.

$[F(2,34) = 41.32, p < .001]$. Bonferroni-corrected paired t-tests confirmed that WD-Name pairs were associated with slower responses than WD-Image or BD pairs (both $p < .001$) but that WD-Image and BD pairs did not significantly differ from each other ($p = .148$).

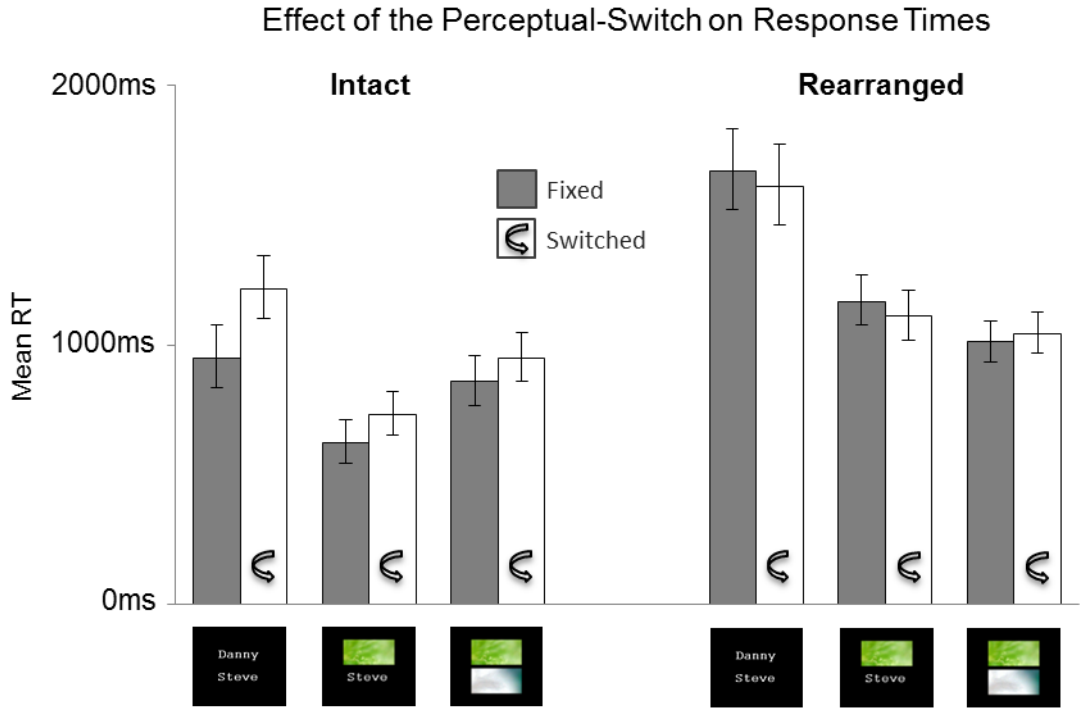


Figure 7.9: Mean reaction times for correct responses, as a function of pair type, test condition and perceptual-switch. Employing the perceptual-switch had the effect of increasing reaction times to intact pairs, but had no effect on rearranged pairs.

Perceptual-switch interacted significantly with test condition $[F(1,17) = 9.96, p = .006]$, reflecting the fact that the perceptual-switch slowed reactions to intact, but not rearranged pairs. Test condition also interacted with pair type $[F(1,17) = 2.68, p = .120]$, reflecting the fact that intact/rearranged differences were smaller for WD-Image pairs than either WD-Name ($p = .014$) or BD ($p < .001$) pairs, which did not significantly differ from each other ($p = .585$).

7.4 Discussion

In broad summary, three main conclusions can be drawn from the results without reference to particular models of recollection and familiarity. First, between-

domain pairs were more easily recognised than within-domain pairs, replicating a result found in other studies in this thesis. Second, switching the position of component items significantly impairs associative discrimination, but only for between-domain pairs. These results suggest that the improved discrimination observed for between-domain pairs relative to within-domain pairs may be at least partly driven by a form of holistic processing, possibly to the extent that these pairs are being perceptually unitized. Finally, however, the perceptual-switch had the effect of increasing reaction times and reducing confidence to intact but not rearranged pairs, and this effect was present across all three pair types. This raises the possibility that all three pair types were recognised holistically to some degree.

7.4.1 Model-specific conclusions

Some model-specific conclusions can also be drawn, though we are cautious about doing so for several reasons. The most important of these is that the parameter estimates, in particular for the DPMSD model, were relatively unstable and sensitive to starting conditions, which is likely to be a consequence of the relatively low number of responses per condition (24). For example, if the DPMSD model is used to estimate the qualitative effect of the perceptual-switch on recollection for *within*-domain pairs (which we did not do since overall performance was unchanged for these pairs), these estimates suggest that perceptually-switching pairs of names or images increased the proportion of rearranged pairs which were recollected, while having no effect on intact pairs. Simultaneously, since performance overall was not affected by the perceptual-switch in these conditions, the strength or diagnosticity of recollection (i.e. the separation of the recollected-intact and recollected-rearranged distributions) was decreased.

How might we interpret such a pattern? We might take it at face value, although it is unclear why the perceptual-switch should make rearranged pairs more likely to be recollected is unclear. Alternatively, the same pattern might arise because of changes in how non-recollected pairs are perceived. In the interests of parsimony, the non-recollected distribution in the DPMSD model is assumed to be equidistant from the recollected-intact and recollected-rearranged distributions. This may not be true in actual fact; for example participants might be more likely to

respond rearranged than intact in the absence of recollection. In such a scenario, where the non-recollected distribution is closer to the recollected-rearranged distribution, some non-recollected trials near the rearranged distribution would be interpreted by the model as being recollected-rearranged (to keep the mean of the non-recollected distribution in the centre, where the model assumes it should be). This would increase the number of trials falling into the recollected-rearranged distribution but also move its mean closer to the recollected-intact distribution, producing the pattern observed here, i.e. higher rates of recollection to rearranged pairs but lower overall recollection strength. Thus, perceptually-switched within-domain pairs might not in fact be recollected any differently, but those which are not recollected may 'feel' less intact as a result of being switched. This would imply that holistic information for these pairs is available and used to make decisions in the absence of recollection; a corollary to this is that such holistic information must not be significantly diagnostic since including a familiarity term in the model does not improve the fit. One possibility is that such decision are based on familiarity of the components, which is greater for non-switched than switched pairs but matched across intact and rearranged pairs, and therefore non-diagnostic for the associative task (Kelley and Wixted, 2001).

In contrast, if conclusions are drawn using a DPSD model, familiarity appears to contribute strongly to associative recognition. Furthermore, the perceptual-switch significantly reduces this familiarity, consistent with the hypothesis that familiarity for associative recognition requires unitization. Disallowing recollection to provide variable information not only worsens the overall fit but also changes the interpretation of the data: it implies that the pairs are being frequently unitized and recognised on the basis of familiarity. Under the better-fitting DPMSD model familiarity does not appear to contribute to discrimination in any of the three conditions. Thus, as in Chapter 5 these results highlight the importance of accurate model selection when drawing conclusions from memory strength data. As discussed previously in Chapter 3, model selection itself should be approached with caution, and preferably supported by independent checks on the assumptions being made in each case. One important strategy is to consider the different conclusions of plausible models alongside each other. In the present case, the conclusions of the two models are strikingly different. Under the assumptions of a DPSD model familiarity can support associative recognition, and is more effec-

tive when to-be-remembered pairs are unitized. Under the assumptions of a finite mixture model familiarity does not contribute to associative recognition, and if unitization supports memory it likely does so by increasing the probability, but not strength, of recollection.

The main way in which the two models differ is in their treatment of recollection. The DPSD model assumes that successful recollection always leads to accurate, high-confidence responses, whereas the mixture model explicitly allows responses based on recollection to vary by describing recollection strength using a Gaussian distribution. Since recollection has previously been shown to produce graded confidence (Mickes et al., 2009, Onyper et al., 2010, Slotnick, 2010), and crucially, accuracy (Chapter 4), the conclusions of the mixture model have a stronger basis in evidence. Corroborating this, the mixture model also provides a better fit to the data here.

Importantly, however, in practice the DPMSD model is rarely used to estimate familiarity and recollection. In several studies, including our own, the DPSD model has been used to support claims that familiarity can support discrimination of novel associations, in particular when such pairs are unitized (Diana et al., 2008, Harlow et al., 2010, Haskins et al., 2008, Yonelinas, 1999). Given that both of these conclusions appear in our own data to be highly model-specific, caution should clearly be applied when such inferences are drawn mainly or exclusively from parameters of the DPSD model. More generally, conclusions should not be made strongly nor presented in isolation when they rely on a single model of memory for which the underlying assumptions have not been thoroughly tested.

7.4.2 A role for unitization in recognition of between-domain pairs?

Regardless of the model used, these results demonstrate that associative recognition of between-domain, but not within-domain pairs, was robustly affected by disrupting the studied spatial relationship between components at test. This result could be seen as consistent with the view that between-domain pairs are encoded in a qualitatively different manner to the within-domain pairs. In particular, they imply that between-domain pairs may have been encoded more holistically, as a single perceptual entity. One potential interpretation of the findings is

that between-domain pairs are more frequently or easily unitized, and therefore rely more heavily on the coherence of the presentation as a whole between study and test.

Unitization has, however, primarily been associated with evidence of familiarity, and the evidence for associative familiarity in this study is model-specific and relatively weak. If unitization does indeed operate for between-domain pairs, the data presented here suggest it is more likely to do so by increasing the success rate of a thresholded recollection process, contrary to the common definition of unitization (Haskins et al., 2008, Quamme et al., 2007). Furthermore, the reaction time data make the picture slightly more complex, since responses to intact pairs of all kinds were slower when they were perceptually-switched, even though this did not significantly affect performance in the case of within-domain pairs.

Perhaps, therefore, qualitatively different information was recollected and used to identify a pair as intact or rearranged. Holistic information might be used to recognise a pair as intact, while explicit retrieval of associations or individual components would be sufficient to identify both intact and rearranged pairs. Although reaction times were increased once the perceptual-switch was introduced, presumably reflecting a reduced availability of holistic information, since there was no reaction time pressure participants may have employed alternative strategies to make a decision in this case. By this explanation, intact pairs of all three pair types could be identified rapidly using holistic information, but between-domain pairs differed in that their recognition was comparatively more dependent on this information; perhaps because it was stronger or more accurate than it was for within-domain pairs. Perhaps such holistic information was also available for within-domain pairs but was in that case not diagnostic of prior occurrence, explaining the fact that the perceptual-switch had an effect on both reaction times and the qualitative pattern of confidence ratings, but a non-significant effect on overall performance. Importantly, (diagnostic) holistic details must have be recollected, since we found no evidence of familiarity. We shall return in more detail to the possibility that associative recognition might be mediated by recollection of holistic details, rather than familiarity for unitized pairs, in the General Discussion (Section 10.5.6).

7.4.3 A role for encoding specificity in recognition of between-domain pairs?

Alternatively, it might be argued that none of the pairs studied were appreciably unitized, and that the interaction between relationship type and perceptual-switch can be explained more simply in terms of Encoding Specificity (Tulving, 1972). Previous studies have highlighted the importance of study-test overlap for successful recognition under different guises (e.g. Transfer-appropriate Processing, (Morris et al., 1977); Context-dependent Memory, (Davies and Thompson, 1988)). If the perceptual study-test overlap is lower for between- than within-domain pairs following the perceptual-switch, this might explain the greater effect of the perceptual-switch on between-domain pairs. Specifically, since names and images are perceptually different, name-image pairs have less vertical symmetry than name-name or image-image pairs. In this case the perceptual-switch might have a correspondingly greater effect on the study-test overlap of between-domain pairs, making them more difficult to recognise. Context-dependent effects have also been shown to be more important for recall than recognition (Godden and Baddeley, 1980), possibly consistent with the recollection-specific effect of perceptual-switch found (and in interesting opposition to the claims about familiarity being the focus of benefits from unitization).

7.4.4 Summary

An explanation based on encoding specificity arguably fits more easily than unitization with the apparent absence of familiarity for associations in this task. Critically however, such an explanation does not readily explain the relationship effect itself, i.e. that between-domain pairs are more readily discriminated in an associative recognition test. Thus overall the results point instead towards an effect which is similar to unitization in some respects - the perceptual-switch disrupted holistic properties of intact pairs, and consequently reduced the ability to recognise them - but is based on recollection and not familiarity. These holistic properties supported recollection across pair types, but between-domain pairs seemed to be more reliant on them, given their particular disruption by the perceptual-switch. Thus perhaps between-domain pairs are recognised more easily than within-domain pairs across these studies for two main reasons. Firstly,

images were more strongly encoded when paired with a name than another image. Secondly, name-image combinations lent themselves better to the perception and encoding of holistic characteristics which later benefitted recollection - perhaps by providing additional cues at retrieval. It may be that this latter effect is particular to the combination of stimuli we have chosen, in the sense that there may exist a framework for attaching lexical labels to unnamed images. These results demonstrate that the relationship between stimuli can have considerable effects on how they are remembered, and hint at some possible mechanisms - but to gain a more complete understanding of how stimuli and their representations interact in memory will require systematic studies using different stimulus combinations and encoding instructions.

Another result from this chapter has been the clear separation of recall-to-accept and recall-to-reject: intact and rearranged pairs were affected very differently by the perceptual-switch. Do recall-to-accept and recall-to-reject simply reflect recollection of different episodes, or are there more qualitative differences between the two? We have established the functional characteristics of recollection, including its separate rate and strength properties, and subsequently highlighted its critical importance in associative recognition. The next step is to move further beyond the umbrella term of recollection and to understand in more depth how different processes might support the recollection of information from the past, including how these processes might differ depending on the availability of cues. One way of achieving this is to examine the neural record for evidence of differential processing of intact and rearranged pairs.

Finally, the interpretation of the results in this chapter depend on whether familiarity did or did not contribute to associative recognition. If the arguments we have made from the behavioural data are flawed, and familiarity does significantly contribute to recognition, then unitization may, after all, be a viable explanation for the pattern of data we have observed. Furthermore, such a finding would call into question the validity of the DPMSD model; as we have argued, one important property of a useful measurement model is that it should produce accurate and testable predictions. Thus in the next chapter we shall use Event-Related Potentials to test the respective, conflicting, predictions from the DPSD and DPMSD models that familiarity does or does not contribute to associative recognition.

Chapter 8

Event-Related Potentials

Event-Related Potentials (ERPs) are a record of the changes in scalp potential correlated with a particular event. They are derived from EEG (the electroencephalogram), a record of electrical potential at the scalp. Despite a proliferation of brain imaging techniques in the past 30 years the electroencephalogram have a number of particular qualities which render it a valuable and unique tool for investigating cognitive processes. Interpretation of the EEG in terms of cognition does, however, require a careful understanding of how the signal is related to brain activity. In this chapter the ERP method is described step-by-step, from the generation of electrical fields by neural activity, to the detection, recording and processing of the electroencephalogram, and the production and interpretation of ERPs. Finally, we outline some key ERP effects of interest for this thesis.

8.1 Neural origin of the electroencephalogram

The electrical activity comprising the electroencephalogram is, in part, originally generated by neural activity. Neurons vary widely in morphology, but almost always consist of a highly branched tree of dendrites (which receive incoming neural or sensory connections), a cell body known as the soma and an axon (which propogates information to neurons or other outputs such as muscle fibres). Neurons form a highly interconnected network via uni-directional junctions known as synapses: a single neuron may receive tens of thousands of such connections

at its dendrites, and in turn project to tens of thousands of other neurons at its axon terminals. Signalling between neurons is mediated by action potentials: short lived spikes in electric potential which propogate rapidly along the axon to synapses at the axon terminal (see for example Lodish et al. 2000 for a detailed description of the action potential). Action potentials transfer information by influencing the probability of a further action potential occurring in each connected neuron.

Neural activity generally refers to this inter-neuron signalling via action potentials. This activity can elicit changes in the electrical field, including at the scalp where it can be non-invasively measured, in two related but distinct ways. Firstly, the action potential directly causes a temporary extracellular electrical field along the length of the axon - inward flow of positive ions at the point of depolarization is balanced by outward flow in passive and recovering regions. This electrical field changes rapidly as the action potential moves along the axon, which typically lasts no more than a millisecond. Before it can fire another action potential, however, the membrane of the neuron must recover towards its resting potential, a process which lasts for tens of milliseconds and engenders an electrical field of the opposite polarity to that produced by the action potential. As a result, the action potentials from a sufficiently large number of similarly oriented neurons will rarely produce large changes in electrical field. This is because at any given moment, the strong electric field changes produced by neurons firing action potentials will be roughly cancelled out by larger numbers of recovering neurons (each producing weaker electric fields in the opposite direction).

Secondly, the axon of a neuron is connected to the dendrites of other neurons via junctions called synapses, which mediate the transfer of information throughout the nervous system. Upon reaching each synapse, an action potential triggers changes in the membrane potential of the post-synaptic neuron. This post-synaptic potential dissipates much more slowly than the action potential and remains relatively localised at the dendrites of the post-synaptic neuron. The effects of multiple pre-synaptic action potentials are thereby summated in the post-synaptic neuron over time according to their synchronicity, synaptic strength, dendrite structure and other factors, and may result in sufficient depolarization to cause an action potential. Depending on the activity of pre-synaptic neurons and the firing rate of the post-synaptic neuron, post-synaptic potentials can last

on the order of hundreds of milliseconds.

Since the postsynaptic potential reflects neural activity of neurons over a similar or longer period than the recovery time for each neuron, it becomes possible for the post-synaptic effects of active neurons to sum together. In contrast to electrical field changes elicited by action potentials, this mechanism allows co-occurrent neural activity to produce large enough changes in electrical field to be detectable at the scalp. The conditions that allow this to happen are discussed in the next section.

8.1.1 Propogation of neural activity to the scalp

Changes in the electrical field will propogate through the brain, and the amplitude of these changes will decrease with distance from the original source. In addition to this, the electrical field does not diffuse evenly in all directions (Wood, 1987). Instead it forms a dipole, in which the electrical field projects strongly along the axis defined by its poles, but weakly in other directions, and not at all perpendicular to this axis. To be detectable, therefore, even a strong dipole must be close to, and oriented towards, a point on the scalp.

In addition to synchronous activity, another determinant of the strength of a dipole is the relative configuration of the active neurons (Allison et al., 1986). When aligned in parallel, so that the positive ends of individual dipoles are oriented in the same direction, the effect of each sums together to create a relatively strong dipole. Such a configuration of similarly oriented neurons is known as an open field; certain configurations of cortical neurons form open fields (Kutas and Dale, 1997). In contrast, when the orientations of each neuron vary, the effects of individual dipoles largely cancel out and cannot be detected at distance. These configurations are known as closed fields; examples include groups of randomly or radially oriented neurons, including neurons in the hippocampus.

In practice, groups of synchronously active neurons will fall between these two extremes, producing correspondingly weak or strong dipoles. Moreover, in other cases the configuration can be activity dependent: for example a group of neurons may be arranged in a closed field when one subset are active, but produce a relatively open field when a different subset are. Similarly, two open fields - even

when they are relatively distal - may be oriented in such a way that they cancel out and produce a smaller effect at the scalp.

The geometric restrictions discussed here have two main consequences for the interpretation of ERP effects. Firstly, only a small - and inconsistent - proportion of neural activity is detectable by EEG. Secondly, it is reasonable to suppose that this detectable activity is produced mainly by pyramidal neurons in the cortex, which are locally interconnected and tend to form open fields close to the scalp. Thus, the activity measured by EEG is likely to reflect activity in cortical circuits, and not in other brain regions such as the hippocampus.

Under certain circumstances, these electrical field changes can manifest as systematic variations in potential at the scalp. While most neural activity gives rise to scalp potential changes only on the order of a few microvolts, these can be measured and recorded using sufficiently sensitive apparatus. The change in this scalp potential over time is known as the Electroencephalogram (EEG).

8.2 Processing the electroencephalogram

Like all functional brain imaging techniques the electroencephalogram is a relatively indirect measure of neural activity, and the raw data must be carefully processed to extract a meaningful signal. The ERP method averages together timelocked segments of data to attenuate the effects of random noise (i.e. background EEG). Thus, ERPs theoretically contain systematic patterns of activity, known as ERP components, related to a cognitive process or task. Before these components can be identified and analysed, however, other noise and artefacts can be compensated for or removed by careful processing.

8.2.1 Forming ERPs from the electroencephalogram

The electroencephalogram can be thought of as comprising a signal from neural activity of interest, together with noise from many sources, including muscle tension or movement, uncorrelated neural activity and the recording equipment. In most cases, these and other sources of noise have a much greater influence on the raw electroencephalogram than the relatively small changes which constitute the

signal of interest (Kutas and Dale, 1997). As a result, the electroencephalogram typically has a low signal to noise ratio (SNR), and only broad, low definition conclusions about neural activity - such as sleep cycles - can be derived directly. Thus, to extract the subtler changes associated with cognitive activity, researchers commonly derive ERPs.

ERPs are the neurally-driven changes in scalp potential correlated with a particular event, such as an experimental trial, constructed from the electroencephalogram by isolating and averaging together segments (epochs) associated with the event being studied. When the event is more strongly correlated with the activity of interest than the sources of noise, averaging in this way attenuates the noise while retaining the signal, improving the SNR. For a perfectly correlated signal, and perfectly uncorrelated noise, the SNR increases as the square root of the number of trials averaged together (Perry, 1966). As a result of this relationship, in modern memory research (including this thesis) a minimum requirement is placed on the number of trials used to form an ERP, commonly 16 trials (Luck, 2005).

8.2.2 Improving ERP signal

In practice, neural activity will rarely be perfectly correlated with the task of interest. For example, participants may simply fail to engage the required cognitive processes on some trials. Temporal differences between trials can also affect the correlation between an event and the activity being studied. Inevitably, the amplitude peak associated with some cognitive process will occur at different times for each trial, making the peak of the averaged signal smaller in amplitude and wider in temporal distribution than the individual trials. Thus, the peak and length of the averaged ERP may differ from the average peak and distribution of the original trials, and should be interpreted with caution (Coles and Rugg, 1995). Conversely, the integral of a voltage deflection over time will not be affected by averaging: the integral of the averaged ERP is equivalent to the true average integral. As a result, ERP components (voltage deflections at the scalp associated with a particular cognitive process or processes) are often characterised by an area-related measure, such as the mean voltage deflection over a particular time interval. In these cases, selecting a time interval which accurately captures

the component of interest is the main challenge, particularly in the presence of distinct but overlapping components.

Conversely, many sources of noise do correlate with the event to some extent and will be attenuated in the averaged ERP to a correspondingly lesser degree. Such systematic sources of noise should ideally be identified, and compensated for, during the design of an experiment or directly removed from the signal after recording. In practice, however, some systematic activity will always remain, and an ERP must be explicitly interpreted as a record of all electrical activity correlated with an event, rather than a pure reflection of the cognitive processes being investigated. Often, the processes being investigated can be targeted more effectively by examining differences between two ERPs, constructed such that the conditions elicit overlapping cognitive requirements except for those under investigation. For example, in order to successfully recognise a previously encountered stimulus a participant will engage myriad processes related to attention, perception, working memory, preparation of motor responses and many others, in addition to those directly supporting recognition in particular. This unwanted activity can be attenuated by comparing the ERP to one elicited by a baseline condition that elicits many of the same supporting processes, such as the correct rejection of a new stimulus, in theory isolating the recognition-related activity of interest.

While averaging can reduce the magnitude of random noise, it cannot remove it completely. As a result ERPs always contain noise, both from systematic sources such as eye movement, and from random sources which are insufficiently reduced by averaging. Careful processing of the signal before or after averaging is therefore employed to remove some of the more common sources of noise. For example, ocular artefacts (such as blinking) can be largely removed by estimating the effect of a given eye movement on each electrode channel and adjusting the signal accordingly (Lins et al., 1993). Voltage drift (a gradual, large scale change in potential across the scalp due to changes in the recording environment) can be reduced by applying a high pass filter, or rejecting epochs for which the voltage changes by more than a predefined value over the course of the epoch. High frequency noise such as muscle activity can be reduced using a low pass filter, trial-by-trial averaging or temporal smoothing, and rejecting epochs contaminated with excessive muscle activity. Segments of the signal containing recording artefacts, or signal

saturation, are also removed prior to processing.

At this point, ERPs can be formed from the remaining processed epochs. As noted above, a minimum number of epochs (16 in this thesis) are normally required for each ERP before it can be included in further analysis, in order to ensure an adequate signal-to-noise ratio. Importantly however, even where an ERP contains a large number of epochs, only noise that is weakly correlated with the event of interest will be attenuated by averaging; any correlated activity, beyond that which has been explicitly removed during processing, will remain. The data must be interpreted carefully, with close reference both to the possible contributing activity and the broader characteristics and limitations of ERPs.

8.3 Interpretation of ERPs

In common with all functional brain imaging methods, ERPs provide only an indirect and limited measure of neural activity and must be interpreted with caution, in reference to these limitations. For example, ERPs and fMRI extract information from completely different physical phenomena (scalp potentials in the case of ERPs; the haemodynamic BOLD response in the case of fMRI), but both techniques indirectly measure task-related changes in activity of neurons summed over time and space, and inferences from either method must be made on this basis.

8.3.1 Spatial and temporal properties of ERPs

Nonetheless, ERPs have particular characteristics and limitations. One main disadvantage of ERPs is that they provide less spatially precise data than some other techniques. This is less an inherent feature of the electrical field itself, from which considerable spatial information could theoretically be extracted, than it is due to the way the electroencephalogram is recorded: at the scalp, after much of the spatial information has been lost. A further (and more fundamental) problem limits the neuroanatomical inferences that can be drawn from scalp potentials. The data collected at the surface of the head are of an inherently lower dimensionality than the set of locations in the brain from which they can

be derived. Thus, for a given distribution of activity at the scalp, an infinite number of sets of neural generators could be responsible: i.e. the inverse problem is ill-posed (Niedermeyer and Lopes da Silva, 2005).

Conversely, an advantage of ERPs is that they uniquely (for non-invasive imaging) measure electrical activity, which is highly responsive to changes in neural activity. In theory, the sampling rate of the digitiser determines the maximum temporal resolution of ERP data. The actual resolution, however, is limited by other factors. For example, the neurally-generated portion of the signal is itself an indirect measure of neural activity, mostly reflecting post-synaptic potentials, which sum pre-synaptic action potentials over tens or hundreds of milliseconds. The temporal resolution of the EEG signal in terms of the original, pre-synaptic neural activity is therefore also limited by the rate at which neurotransmitters are released and transferred across the synapse.

Moreover, the onset of an ERP difference provides only an upper bound on the onset of its associated cognitive process, since it may reflect processing downstream from earlier, less visible differences in neural activity. Similarly, the temporal properties of an ERP component reflect those of an average of several individual trials. Any onset or offset differences between trials will translate as an increase in the length, and corresponding reduction in amplitude, of the averaged component. An ERP correlate may therefore be shorter (due to undetected activity) or longer (due to averaging) than the cognitive process it reflects.

Furthermore, in order to reduce the effects of short-lived sampling artefacts, ERPs may be smoothed before analysis - data points are adjusted towards a weighted average of its temporal neighbours. This will reduce the true temporal resolution of the smoothed ERP by a factor determined by the size and shape of the filter used. Nonetheless, ERPs retain a temporal resolution of around 10-100Hz (i.e. they are precise to tens of milliseconds), which is on a similar order to the rate at which information is transferred between neurons. Thus changes in the ERP are highly responsive to changes in neural activity, and several orders of magnitude more responsive than functional imaging techniques which measure haemodynamic changes, such as fMRI (which is typically precise on the order of seconds) or PET (minutes). ERPs therefore have a relative advantage in terms of temporal resolution over most other functional imaging techniques.

Finally, as with all existing imaging methods, ERPs are a highly selective measure of neural activity (see Section 8.1.1). The interpretation of ERP components is primarily restricted to activity in clusters of relatively synchronous pyramidal cortical neurons, arranged in open fields close to the scalp. Thus, an ERP effect should not be interpreted as a record of all brain activity for a given process. This leads to an important point: null results in ERP studies (e.g. matched scalp activity across two conditions) do not imply that the underlying neural activity is identical, simply that the activity in the subset of detectable neurons might be (Coles and Rugg, 1995, Kutas and Dale, 1997). Even if the activity of neural generators is identical across two conditions, one cannot draw similar conclusions about the remaining, overwhelmingly large subset of the brain from which little or no activity can be detected. While this may at first appear to be a major weakness of ERPs, it is important to bear in mind that similar restrictions apply to all imaging techniques. Moreover, limiting observed activity to particular subsets of neurons may in some cases have advantages as well as drawbacks. Restricting the observed activity to a subset of the brain reduces the number of components contributing to ERPs, potentially making it easier to isolate and manipulate those which do appear robustly across experiments.

8.3.2 ERPs and cognition

The selectivity described above means that ERPs are well suited to demonstrating differences between conditions. Similarities between conditions in exploratory studies are generally uninformative, although under well-defined hypotheses the absence of specific, expected differences in particular components can be interpreted as providing some evidence for specific functional overlap. It is also important to bear in mind that the neural activity detected is inherently correlational, in other words it does not necessarily reflect the direct manifestation of the cognitive processes it is elicited with. Nevertheless, the dependence of ERPs on cortical networks and their high temporal resolution mean that, under certain assumptions, ERPs can be a valuable tool for investigating cognitive phenomena in particular.

The main assumption required to draw cognitive inferences from ERPs is the idea that there an equivalence between cognitive processes and patterns of neural

activation: differences in one implies differences in the other. On this basis, differences in scalp activity across conditions are assumed to reflect some cognitive difference, whose function can be deduced by careful design of the conditions being compared.

One way of isolating this function is by using difference ERPs, as suggested in Section 8.2.1. Here a cognitive process is carefully isolated by examining the difference between ERPs for two carefully chosen conditions and relating it to the functional difference between them. This approach also makes the assumption that cognitive processes are additive and do not interact, i.e. the pure insertion principle (Donders, 1868, Sternberg, 1969). In practice some violation of this assumption is likely (Friston et al., 1996), for example the latency of a shared process might change when additional processes are engaged in one condition, therefore caution should be taken not to abscribe process-purity even to difference ERPs. Violation of the pure insertion principle is not unique to ERP research however, and so long as the appropriate caveats are observed meaningful cognitive information can still be inferred.

8.3.3 Inferences from ERPs

Differences between ERPs can be characterised in several ways, with distinct cognitive interpretations. Firstly, when the spatial and temporal distribution of an effect are matched across two conditions but differ in magnitude, this may imply that the same cognitive process is engaged in both conditions, but to a greater degree (or on more trials) in one than the other (the first inference). The strength of this form of inference is, however, dependent on the way in which amplitudes are measured. Peak amplitude differences are extremely sensitive to latency jitter (see Section 8.2.1); area measures are less susceptible but still may be affected to some degree (Handy, 2005). Therefore magnitude differences should be treated with caution when there is also evidence of temporal differences.

Secondly, latency differences between components with similar spatial distributions might indicate similar cognitive processes are engaged earlier in one condition than another (the second inference). Caveats apply to conclusions drawn from ERP dynamics (see Section 8.3.1) and temporal differences between ERPs cannot normally be straightforwardly interpreted as equivalent temporal differ-

ences in the cognitive processes (Coles and Rugg, 1995). Nonetheless, genuine differences in ERP dynamics can still indicate that different neural populations are engaged.

Thirdly, differences in the spatial distribution (topography) of an effect are normally also taken to indicate the engagement of non-overlapping neural populations (the third inference), under the assumption of cognitive and neural equivalence stated above. Whether this reflects the engagement of a neuroanatomically distinct set of generators or simply differential relative engagement of existing ones is difficult to determine without employing source localisation.

Where differences exist between ERPs for two conditions of interest, these can take the form of quantitative (magnitude) or qualitative (topographic or latency) differences, each associated with particular cognitive inferences. Accurate characterisation of the differences between two conditions is therefore an important part of the ERP method.

8.3.4 Analysing ERPs

ERP differences are normally confirmed and characterised using inferential statistics, which is the approach taken in this thesis. The reliability of overall magnitude differences between ERPs can be assessed using a suitable repeated measures ANOVA, and applying a correction for the violation of sphericity in ERP data. In all of the ERP analyses which follow in this thesis we use the Greenhouse-Geisser correction whenever the assumption of sphericity is violated (Jennings and Wood, 1976). Topographically relevant factors (such as hemisphere, superior/inferior sites and frontal/parietal location) are often included to allow better characterisation of effects, and guide follow-up tests. Alternatively, a component of interest can be identified in advance, and magnitude differences for the relevant electrodes and time window are then compared across conditions.

If magnitude differences are present, it is important to assess whether they arise because of greater engagement of equivalent cognitive processes (the first inference in 8.3.3) or topographic differences caused by the engagement of distinct cognitive processes (the third inference). This is not straightforward, since ERP data is multiplicative (the activity of a neural generator changes voltage unevenly across electrodes) while the ANOVA model is additive (it assumes a constant change

at each electrode). Thus magnitude changes in a single generator can introduce differences in spatial distribution at the scalp, which are interpreted by the model as interactions between effect and location, indistinguishable from the engagement of different neural generators.

To account for this, prior to topographic analysis the ERP data can be rescaled in such a way that the absolute voltage is matched across conditions, but the relative pattern of activities are preserved. One (widely used) such rescaling is the max/min method (McCarthy and Wood, 1985), though other approaches have been suggested, e.g. vector normalization (Glaser and Ruchkin, 1976). There exists considerable debate over the validity of conclusions drawn from rescaled data. For example, some authors suggest that the use of rescaling can lead to an increase in type II errors in the case of the max/min method, or type I errors in the case of vector normalization (Urbach and Kutas, 2002). Others however argue that rescaled data can be validly used to check for the *existence* of distribution differences, although these differences should be *characterised* by reference to the original, non-rescaled data (Wilding, 2006). This is the approach taken in this thesis, and data are rescaled using the max/min method. The methods used to record and process the ERP data in this thesis are briefly described in the next section.

8.4 ERP correlates of episodic retrieval

In this thesis we are concerned primarily with ERP effects that relate to familiarity and recollection. Old/new effects have been widely studied in recognition memory, and we outline here some effects which we might expect to be informative. This is by no means intended to be an exhaustive review of the ERP literature, however, even within the focused realm of episodic old/new effects at retrieval. For that purpose we direct the reader to a variety of more comprehensive reviews of this field (e.g. Curran et al., 2006, Donaldson et al., 2002, Rugg and Allan, 2000) and note also that the interpretations of ERP correlates (and by extension, the conclusions of studies such as this which use these interpretations) are constantly evolving on the basis of new evidence.

8.4.1 The FN400 effect

In the experiments which follow, we examine old/new difference ERPs at retrieval, conditional on whether the old items were presented as part of an intact or rearranged pair. If old and new pairs are distinguished on the basis of familiarity, both intact and rearranged pairs should show evidence of the FN400 effect, a positive deflection relative to new pairs at mid-frontal electrodes around 300-500ms post-stimulus (Rugg et al., 1998). The FN400 has been linked to familiarity on the basis of its correlation with putative behavioural measures of familiarity (e.g. Curran, 1999; 2000, Curran and Cleary, 2003, Nessler et al., 2001, Trott et al., 1999; see Rugg and Curran, 2007 for a review). Nonetheless, the mapping is not universally accepted - for example, the FN400 has also been interpreted by some as reflecting conceptual priming rather than familiarity (Olichney et al., 2000, Yovel and Paller, 2004).

Given that priming may also provide a means of distinguishing old from new stimuli, evidence of familiarity in the behavioural data does not explicitly differentiate between the effects of familiarity and those of priming. For example, if priming provides some evidence of previous experience for every old pair, and this evidence is generally weaker than that resulting from successful episodic recollection, then priming and familiarity would be indistinguishable using confidence data. Nevertheless, ascribing the DPMSD estimate of familiarity to priming effects would lead to the same predictions for the imaging data, i.e. an FN400 effect for both intact and rearranged pairs relative to new pairs. Crucially, intact and rearranged pairs should not differ in the size of the FN400. If they did, this would imply that the familiarity signal (or strength of evidence from priming) differed according to whether a pair was intact or rearranged, and the signal would therefore be diagnostic of this difference. The behavioural evidence so far indicates that in fact familiarity provides at best, non-significant evidence that a pair is intact or rearranged. If the DPMSD model is correct and this is the case, the FN400 should be matched across intact and rearranged pairs.

8.4.2 The LPONE

Secondly, since the model estimates indicate that recollection occurs for both intact and rearranged pairs, we might expect to see evidence of the left parietal old/new effect (LPONE), a positive deflection relative to new items at left parietal electrodes between 500-800ms post-stimulus. The LPONE has been associated with recollection, especially for studies using words as stimuli (Paller and Kutas, 1992, Rugg and Yonelinas, 2003, Smith, 1993), and is often considered to be one of the most reliable electrophysiological old/new effects. Others have found evidence, however, that recollection for other stimulus types such as faces may elicit differently distributed effects (MacKenzie and Donaldson, 2007, Yick and Wilding, 2008, Yovel and Paller, 2004), or that the distribution of the LPONE may depend on the presence of lexical or pictorial material (Galli and Otten, 2011). In a related finding, the presence of the left parietal old/new effect was linked to the meaningfulness of stimuli (Cycowicz and Friedman, 2007), such that for meaningless stimuli the LPONE was absent and instead a weaker, centrally located positivity was observed during the same time window. An interesting question for the following study, therefore, is whether names and abstract images differ in terms of the distribution of the LPONE (or indeed, whether the LPONE is present for abstract images at all).

8.4.3 The right-frontal effect and the LPN

Some later ERP effects, occurring subsequent to the LPONE, have also been associated with old/new recognition tasks. The right-frontal effect is characterised by a positive deflection for old relative to new stimuli at inferior frontal electrodes on the right of the scalp, and has been linked to post-retrieval processing, such as source monitoring (Mecklinger, 2000, Ranganath and Paller, 2000, Wilding, 1999). Finally, the late posterior negativity (LPN), a negative deflection for old relative to new items with a superior parietal focus, has often been observed to follow the LPONE. The exact interpretation of this effect is a matter of some debate. The LPN has been linked to the reconstruction and integration of episodic information (Cycowicz and Friedman, 2003, Cycowicz et al., 2001, Johansson and Mecklinger, 2003, Johansson et al., 2002, Li et al., 2004, Wolk et al., 2007). In this case, we should expect the LPN to be more pronounced for rearranged pairs,

which may prompt recollection of two separate episodes, compared to intact pairs, which are linked to just a single episode.

Some have also suggested that the LPN reflects perceptual retrieval in particular, given its location over occipital electrodes and association with visual stimuli (Wolk et al., 2007). If this is the case, rearranged pairs might again show a greater effect, but this should interact with the type of pair. The images in our studies contain considerably more perceptual information than names do, and so the size of the LPN should track with the number of images in each pair.

Finally, others have suggested that the LPN may in fact reflect differences in response preparation, i.e. it is a manipulation of the Lateralized Readiness Potential (Kuo and Van Petten, 2006). In this case, the lateral position of the LPN should vary with response hand, and be present over the opposite hemisphere to the hand being used to respond. We can discriminate between these three interpretations of the LPN in the following experiment, by testing 1) whether rearranged pairs elicit larger LPNs than intact pairs; 2) whether abstract images elicit larger LPNs than names and 3) whether the response hand used influences the lateral position of the LPN.

8.5 ERPs in this Thesis

The EEG data in this thesis were recorded from 64 silver/silver-chloride electrodes, embedded in an elasticated Quick-Cap (Neuromedical Supplies; www.neuro.com). Electrode locations were a subset of those defined in the extended 10/20 system (Chatrian et al., 1985): FPZ, FP1, FP2, AF3, AF4, FZ, F1, F2, F3, F4, F5, F6, F7, F8, FCZ, FC1, FC2, FC3, FC4, FC5, FC6, FT7, FT8, CZ, C1, C2, C3, C4, C5, C6, T7, T8, CPZ, CP1, CP2, CP3, CP4, CP5, CP6, TP7, TP8, PZ, P1, P2, P3, P4, P5, P6, P7, P8, POZ, PO1, PO2, PO3, PO4, PO5, PO6, PO7, PO8, CB1, CB2, OZ, O1, O2. The ground electrode GND was positioned frontally, midway between AF3 and AF4. During recording, data from each electrode channel were referenced to an additional electrode located midway between CZ and CPZ. All channels were re-referenced offline to a virtual midline, calculated by averaging the signal from two electrodes M1 and M2 located on the left and right mastoids. Data were also recorded from two pairs of linked electrooculogram (EOG) elec-

trodes: VEOU and VEOL positioned above and below the left eye; HEOL and HEOR positioned on the outer canthi.

Before recording, the cap was carefully fitted and secured, and the impedance at each electrode was brought to below $2k\Omega$ by inserting a conductive gel between the electrode surface and scalp, and gently abrading the skin as necessary. Participants were seated comfortably and spent some time observing and manipulating the effects of eye movement, blinking, muscle tension and other movement on the EEG signal, until they could comfortably minimise the associated noise.

During recording, each channel was amplified and band-pass filtered between 0.01Hz and 40Hz. The data were digitized by a 16 bit analogue to digital converter at a sampling rate of 250Hz and recorded to a desktop computer using Neuroscan Acquire (version 4.3, www.neuroscan.com). The experiment was run using EPrime (version 1.1, www.pst.net) on a separate desktop computer. The EEG recording was timestamped when stimuli were presented, so that ERPs could later be formed from epochs timelocked to the onset of each stimulus presentation.

Once recording was completed, the EEG was visually inspected to remove segments contaminated with muscle activity or excessive horizontal eye movement (by inspecting the HEOG channel), and then processed using Neuroscan Edit (version 4.3). For each participant at least 32 epochs containing blinks (but no significant muscle tension or horizontal eye movement) were chosen and used to compensate for vertical eye movement by regressing each active electrode against the VEOG channel.

Epochs of interest were timelocked around stimulus presentations. Each epoch was baseline corrected by subtracting the average activity at each electrode over a 148ms pre-stimulus interval, and then smoothed using a uniform distribution filter. A drift detection algorithm was applied, which identified and removed any epochs for which one or more active electrodes varied in amplitude by more than 75μV between the first and last data points, a total period of approximately 2000ms. Epochs containing one or more active electrode exceeding $\pm 100\mu$V after baseline corrrection and smoothing were also removed. Finally, ERPs were formed for each condition of interest by averaging together relevant epochs from those that remained. Participants with fewer than sixteen trials in any condition of

interest were rejected prior to analysis of the ERPs.

8.6 Summary

The relationships between cognition and neural activity, and between neural activity and ERPs, are complex and dictate the inferences that can be made about cognition using ERPs. However, when properly understood and carefully interpreted, ERPs provide unique functional imaging data which complement the very different characteristics of other imaging modalities such as fMRI or PET. This is particularly true when the structure and location of networks underlying a process are of less interest than the dynamics or function of that process itself. The consistency of findings across the ERP literature, and the replication of well-defined components, support the use of ERPs as a powerful tool for investigating cognitive and psychological theories.

Chapter 9

ERPs in Associative Recognition

9.1 Introduction

So far in this thesis we have reported a number of behavioural studies, which go some way to characterizing the phenomena of recollection and familiarity that underpin episodic memory. Briefly, the two are functionally separable: while familiarity is a continuous signal, recollection is a thresholded (but graded) phenomenon which can succeed or fail (Chapter 4). Associative recognition depends on successful recollection (Chapters 5–6), while familiarity is diagnostic only of item recognition (Chapter 6).

9.1.1 Testing the DPMSD model predictions

The first conclusion, characterizing recollection as thresholded and graded, and familiarity as continuous, suggests the use of a particular quantitative memory model (the DPMSD model), and the rejection of others (such as the more commonly used DPSD and UVSD models). The second conclusion, that familiarity is non-diagnostic for associative recognition and recollection is critical, depends on this DPMSD model. The model's assumptions are supported by the data in Chapter 4 and it provides a good fit to the behavioural data reported in this thesis, but before accepting a conclusion based primarily on its parameter estimates it is equally important to test the model by verifying its predictions in other domains. Here we test the prediction of the model that recollection, but not

familiarity, contributes to the discrimination of intact from rearranged pairs by examining Event-Related Potentials (ERPs) elicited by these response categories. As we have outlined in the previous chapter, both recollection and familiarity are associated with distinctive correlates in the neural record, taking the form of positive deflections to previously-studied stimuli compared to unstudied lures. The ERP record can therefore be used to examine the contribution of recollection and familiarity to task performance.

9.1.2 Dissociating recollection using ERPs

We can also use ERPs to further dissociate some of the processes which underpin recollection. While we have demonstrated that associative recognition for the stimulus pairs used in this thesis is mediated by recollection, the frequency of this recollection depends, in a relatively complex way, on the relationship between the items being associated (Chapter 5). In particular, the evidence suggests that pairs of dissimilar (between-domain) items are recognised more easily than similar (within-domain) pairs, primarily reflecting an increased rate of recollection. The current chapter asks whether we can further isolate how this advantage occurs by finding differences in the ERPs to each pair type, or test condition. For example, the processes underlying recollection might differ according to the type of information being recollected, or the information retrieved might be interrogated in a different way. These possibilities should be reflected by differences in ERPs linked to recollection or post-retrieval monitoring respectively. Alternatively, the processes comprising recollection may be qualitatively identical.

In Chapter 7 we investigated whether some pairs of un-associated items might be 'unitized', that is perceived in such a way that they may be encoded and retrieved as a single item, allowing their recognition to be mediated by familiarity. We found some evidence which was consistent with unitization, but crucially we also found that the predicted contribution of familiarity - according to the DPMSD model - was absent. Instead, we theorised that while intact pairs might be sometimes recognised on the basis of holistic characteristics, this had the effect of increasing the rate of recollection, and not of allowing familiarity to contribute, as unitization predicts. This would mean that even when holistic information was diagnostic, associative recognition would still be functionally distinct from

item recognition in terms of the memory signals capable of supporting it. How, then, might this be reflected in the neural record? If unitization does improve recognition for some pairs via familiarity, evidence of this should be present in the form of larger familiarity-linked ERP effects to intact than rearranged pairs, though perhaps only for between-domain pairs since these showed the greatest evidence of unitization. Alternatively, if recollection of intact pairs is sometimes based on holistic features, retrieval of this type of information might be reflected in intact/rearranged ERP differences, given that such features should be absent for rearranged pairs. In this latter case, however, these effects should be distinct from those which normally accompany item familiarity. In the current chapter we aim to distinguish between these accounts by comparing ERPs to intact and rearranged pairs.

9.1.3 Interpreting ERP data in cognitive terms

While finding differences between ERPs to intact and rearranged pairs can help us to determine whether neural activity differs across these tasks, it does not necessarily allow us to infer what these differences reflect, and we should be cautious about doing so without reference to the wider ERP literature. An advantage of imaging studies is the multidimensionality and richness of their resulting datasets, which allow subtle differences across conditions to be observed. The large number of possible relationships between aspects of such a large dataset and the phenomena being investigated does, however, lead inevitably to a greater number of statistically significant results arising through noise. This means that results have to be interpreted carefully to reduce the risk of accepting statistically 'significant' relationships which do not actually reflect interesting or repeatable phenomena. In particular, hypotheses about ERP data should be focused and specific, and grounded in theories about what cognitive processes particular neural correlates reflect, and therefore how they should vary within a given contrast.

To mitigate the problem of finding spurious significant effects, and also to place the results of the analysis into a cognitive context, we shall interpret the ERP data presented here primarily in terms of the widely-reported and consistent old/new effects described in the previous chapter: The FN400, the LPONE and the LPN.

9.1.4 Comparing intact and rearranged ERPs directly

Since the primary aim of the current experiment was to examine the neural basis of the discrimination of intact from rearranged pairs, we also directly compared ERPs to intact and rearranged pairs. As well as examining this contrast in the current experiment, we also analysed electrophysiological data gathered during Chapter 5, Experiment 1. No new items were used at test in this earlier experiment, and as a result the primary task being performed by participants while EEG was recorded was the discrimination of intact from rearranged pairs. It is harder to make this claim strongly for the new experiment introduced in this chapter, since participants made a 3-way intact/rearranged/new decision. As a result, the requirement to additionally identify items as old or new brings greater uncertainty to the particular cognitive task being performed at any one time, such as during a specific epoch.

Employing new items does, however, allow comparison with a large number of recognition studies in which the ERP correlates outlined above have been identified and linked to particular memory-related processes. A further important question for this chapter, therefore, is how appropriate the introduction of new items as a baseline is for imaging experiments which aim to investigate associative recognition. In this chapter we consider both approaches, and in doing so insulate ourselves to some extent from this problem. This also, however, provides us scope to investigate the question by comparing behavioural and electrophysiological data from the two approaches. We may expect to observe some differences between the two experiments as a result.

9.1.5 Aims

In this chapter we shall examine ERPs to intact and rearranged pairs from Experiment 1, Chapter 5. To aid interpretation of any differences, we also examine ERPs to intact, rearranged and entirely new pairs from 24 further participants. With these data we aim to answer several questions. First, does the inclusion of new pairs change the difference between ERPs to intact and rearranged pairs? Second, does the neural record contain evidence of greater familiarity for intact than rearranged associations, or does it instead corroborate the conclusion from

the DPMSD model that familiarity does not significantly support associative recognition? Third, do we find evidence of the LPONE, and if so is it material-specific? Fourth, do we find evidence of the LPN and if so, does it reflect episodic reconstruction, perceptual retrieval or simply response preparation?

9.2 Methods

We collected continuous electrophysiological data from participants in two experiments. The first of these, requiring participants to distinguish intact from rearranged pairs, was originally reported in Chapter 5 and the appropriate methods can be found in that chapter. The second experiment required participants to discriminate intact, rearranged and entirely new pairs, and is reported here for the first time. This section describes the participants and procedure for this latter experiment. The EEG recording procedure was identical for both experiments and is detailed in Chapter 8.

9.2.1 Participants

A total of 37 right-handed, native English speakers participated in the experiment. Thirteen participants were excluded from the analysis because they produced fewer than 16 correct, artifact-free responses in at least one condition and of the remaining 24 participants 17 were female (mean age 20.1, range 18-25).

9.2.2 Stimuli

Each stimulus comprised a pair of items presented above and below central fixation. We employed the three stimulus conditions used in the previous chapters. Within-domain conditions comprised pairs of either Christian names (WD-Names) or abstract images (WD-Images); a between-domain (BD) condition comprised equal proportions of image-name and name-image pairs (Figure 3.2). In total 480 names and 480 images were used, see Chapter 3 for details of the stimulus properties and how they were selected. At test, participants were shown either intact (components appeared together at study), rearranged (components

appeared in separate study pairs) or new pairs (neither component appeared at study), Figure 9.1.

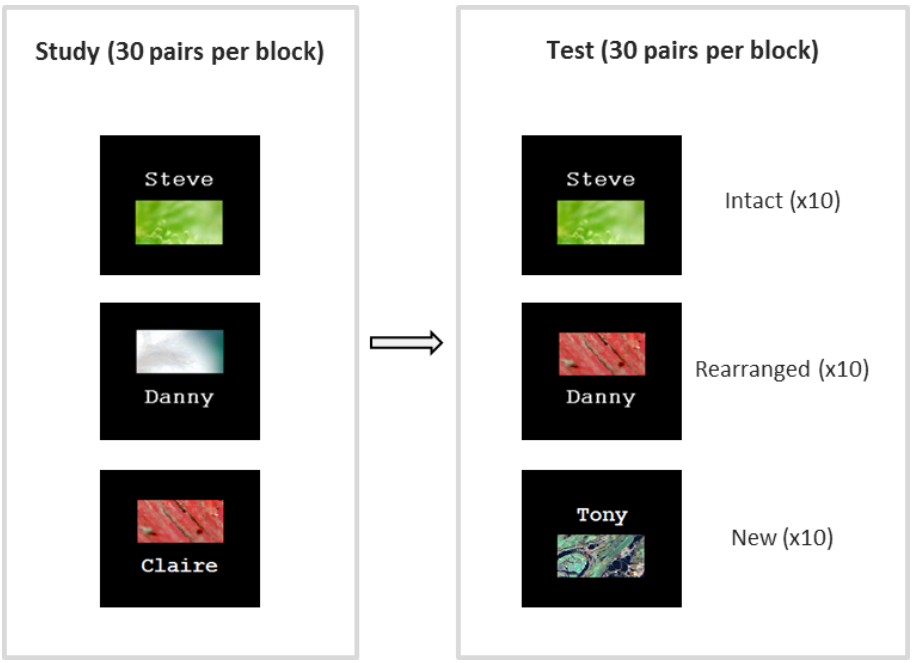


Figure 9.1: The associative recognition task. The experiment was similar to those previously reported in this thesis, with the exception that pairs of entirely new items were presented at test to provide a baseline for examination of old/new ERP effects.

9.2.3 Procedure

Detailed aspects of the experimental procedure not included here can be found in Section 3.1. The experiment was divided into 12 blocks, 4 for each relationship condition (*WD-Names, WD-Images, BD*), and each block was further subdivided into a study phase of 30 trials followed by a test phase of 30 trials. Each study trial contributed to exactly one trial at test: 10 pairs were later shown intact and 20 components (one from each of the remaining 20 study pairs) were used to form 10 rearranged test pairs. Finally, 10 pairs of entirely new items were shown at test.

The study and test procedures are illustrated in Figure 9.2. Each study trial was preceded by a 1000ms fixation cross, and consisted of a pair of items presented for 3000ms above and below central fixation. After each study presentation participants were required to indicate on a scale from 1-5 how well the two items

went together; this response initiated the beginning of the next study trial.

At test, following a 1000ms fixation cross, participants were presented with a pair of items for 1000ms above and below central fixation (associative recognition presentation). Each pair was judged intact, rearranged, or new using the buttons 1, 3 and 5 on the button box (order reversed across half of participants). After a 500ms blank screen participants indicated how confident they were that they were correct, on a scale of 1-5. This confidence response initiated the beginning of the next test trial. Trials were randomly intermixed at study and test and in total each participant completed 40 intact, 40 rearranged and 40 new trials per stimulus condition.

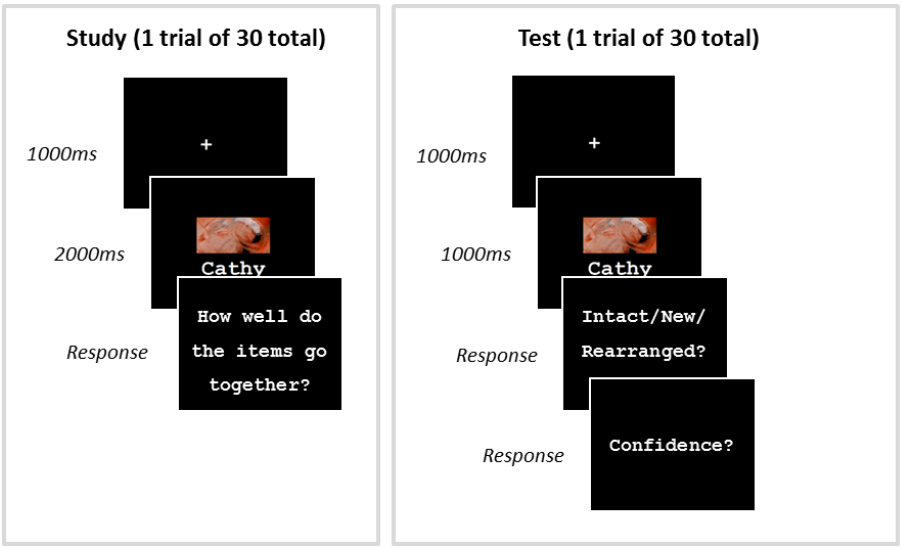


Figure 9.2: Study and test procedures. Every screen was followed by a blank (black) screen for 500ms (not shown in the figure), with the exception of the fixation cross, which was followed by a 100ms blank screen, and the test presentation, which was followed by a 1000ms blank screen.

At both study and test the mapping of left and right buttons to (*intact*/*new*) and (1-5) responses was identically counterbalanced across blocks of 4 participants for each group. The stimulus condition (3: *WD*/*WD*/*BD*) and test condition (4: *intact*/*rearranged*/*new*/*not shown*) of each stimulus component (i.e. each individual word or image) was fully counterbalanced across participants. On average the procedure took 2 hours to complete, including a practice block and debriefing, plus on average one hour spent preparing for EEG recording.

9.3 Behavioural Results

Behavioural results for Experiment 1 are detailed in Chapter 5. Two participants were excluded from the ERP analysis, the data for the remaining 27 was re-analysed and did not differ qualitatively from the full dataset. All behavioural results from Chapter 5, Experiment 1 reported in this chapter are derived from this reduced set of 27 participants.

We assessed old/new and intact/rearranged discrimination in Experiment 2 using d_a. Unlike previous experiments in this thesis, participants made both old/new and intact/rearranged decisions with a single intact/rearranged/new judgment; here we clarify how each response was incorporated into the two sets of confidence data. For the old/new decision, intact and rearranged were both considered 'old', thus an intact pair reported 'rearranged' would be a correct old response. For the intact/rearranged decision, to most closely approximate the data from previous chapters (where all pairs comprised old items) we discarded new stimuli but also 'new' responses, only including old stimuli which were recognised as old. It should be noted that resulting performance estimates will be slightly exaggerated, since the number of (incorrect) low strength or guessed trials is reduced by discarding new responses. Nonetheless, after removing 'new' responses the qualitative pattern in the confidence data should reflect more closely the intact/rearranged decision specifically.

The results (Figure 9.3) showed a similar pattern of performance to that observed in previous chapters. Separate repeated measures ANOVAs revealed a significant effect of pair type on both old/new [$F(1.49, 34.23) = 58.81, p < .001$] and intact/rearranged discrimination [$F(2, 46) = 39.79, p < .001$]. Paired t-tests revealed that in each case the difference was driven by lower discrimination for WD-Image than BD or WD-Name pairs (all $p < .001$) and not by differences between BD and WD-Name pairs (both $p > .999$).

We used linear regression to assess whether item type (0 *names*; 1 *name*; 2 *names*) and relationship type (*within-domain*; *between-domain*) contributed to old/new (Table 9.1) and intact/rearranged (Table 9.2) discrimination. Consistent with previous analyses (see Sections 5.2.2, 6.3 and 7.3), we found significant effects on performance from both item (*names* > *images*) and relationship type (*between-domain* > *within-domain*), for both old/new and intact/rearranged dis-

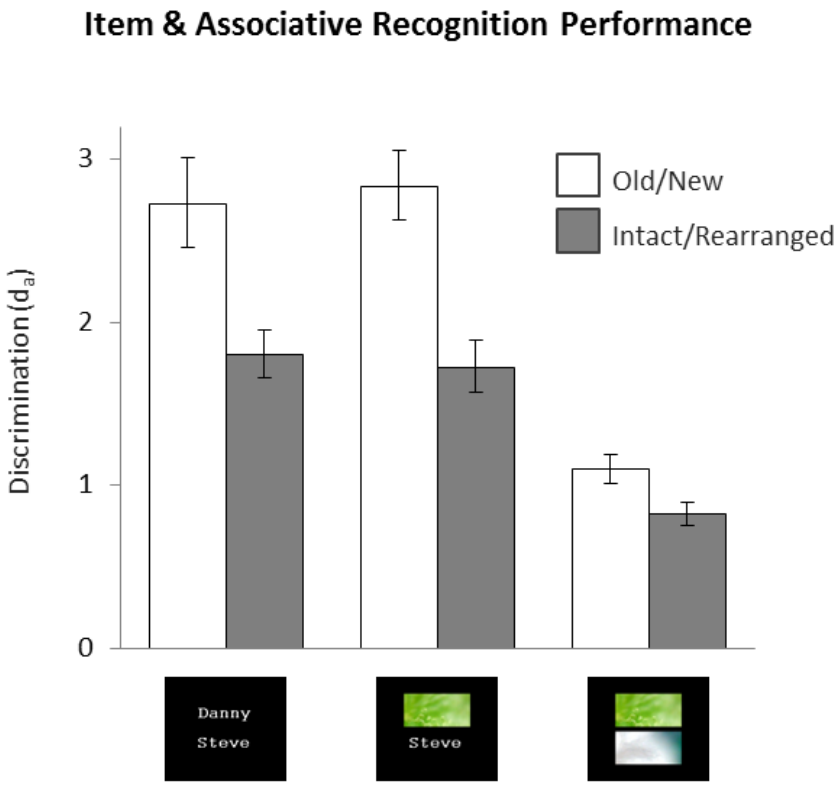


Figure 9.3: Summary of old/new and intact/rearranged discrimination. Overall performance for WD-Name and BD pairs were closely matched for both types, though this belied some qualitative differences in both reaction times and confidence responses.

crimination. Thus, despite introducing pairs of new items as a baseline for the interpretation of ERPs, the behavioural data in Experiment 2 shows the crucial behavioural pattern observed previously: old/new and intact/rearranged discrimination is better when pairs are between- than within-domain.

Parameter	Direction	B	S.E.	β	t	p
(Constant)		1.173	0.245		4.781	<.001
Item type	Names > Images	0.896	0.301	0.287	2.982	.004
Relationship type	BD > WD	1.898	0.347	0.527	5.470	<.001

Table 9.1: Linear regression factors contributing to old/new discrimination. Item (names > images) and relationship type (BD > WD) predict old/new discrimination.

Parameter	Direction	B	S.E.	β	t	p
(Constant)		0.894	0.148		6.026	<.001
Item type	Names > Images	1.046	0.210	0.496	4.988	<.001
Relationship type	BD > WD	0.484	0.182	0.265	2.665	.010

Table 9.2: Linear regression factors contributing to associative discrimination. Item (names > images) and relationship type (BD > WD) predict intact/rearranged discrimination.

As is clear from Figure 9.3, both old/new and intact/rearranged discrimination were more difficult for WD-Image than WD-Name or BD pairs. These overall performance differences were, however, accompanied by some qualitative differences across pair types. We investigated the roles of recollection and familiarity in old/new and intact/rearranged discrimination by fitting each participant's confidence data to the DPMSD model of recognition.

9.3.1 Qualitative old/new differences across pair types

We first checked the significance of the familiarity parameter d'_F in the old/new data using likelihood ratio tests. In contrast to the component recognition data analysed in Chapter 6, including familiarity did not improve the fit of the DPMSD model [$\chi^2(108) = 51.27, p = .969$]. Was this because the task relied solely on recollection? We believe this is unlikely, since when familiarity was excluded from the model, estimates of recollection rate approached ceiling (mean $p(R_a) = .97$). Thus there were not significant numbers of information-free guesses, as there would likely be if (thresholded) recollection was necessary for recognition. Instead, it seems more likely that target distribution is simply too complex to be split into just two normal distributions which reflect the presence and absence of recollection. There are several reasons for this, but foremost is that a greater number of sources (and therefore combinations of sources) of memory evidence are available for this old/new task than the component recognition task. In the component recognition task, a single item is shown at test and judged old or new on the basis of recollection or familiarity. In the old/new task in this chapter targets instead comprise pairs of items, for which either, both, or neither of two components, some originating from different episodes and others not, may be recollected. Additionally, each component also provides some familiarity. As we have demonstrated in Section 3.2, even simple simulated data from the DPMSD model is often more parsimoniously described (though, importantly, not explained) by a UVSD model. Given so many possible combinations of evidence in this task, at best the target distribution will consist of many more than two peaks, and in practice will closely approximate a single normal distribution, making it virtually impossible to estimate the separate contributions of familiarity and recollection.

In theory, recollection rates might to some extent be inferred using estimates from the intact/rearranged data, under the assumption that recollection which supports associative recognition should usually support recognition of individual components. The relationship between recollection of associative and component information is likely to be quite complex however, and the results from Chapter 6 (illustrated in Figure 6.9) additionally suggest that this relationship may interact with component type.

Instead therefore, we restrict our examination of the old/new data here to dif-

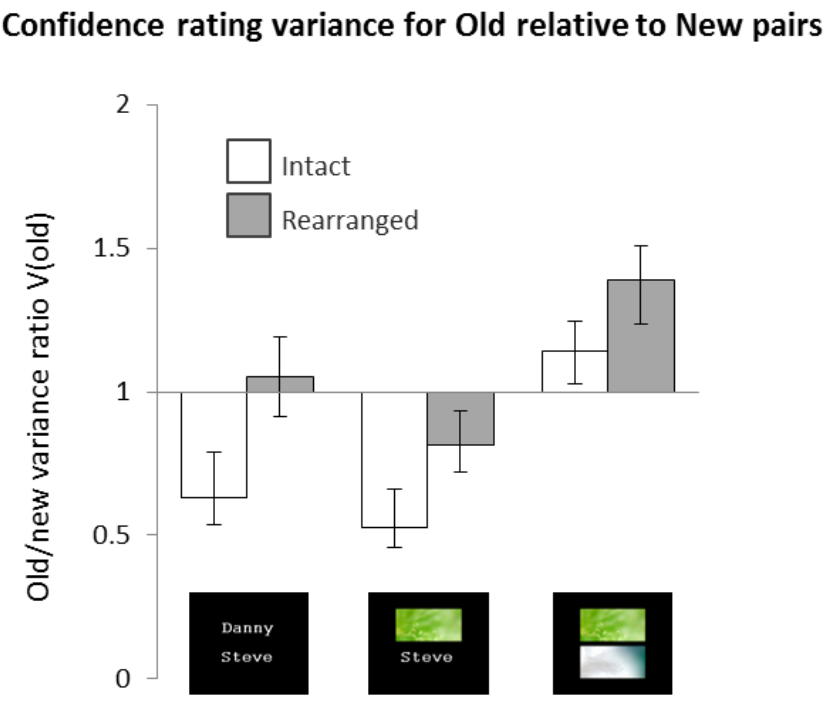


Figure 9.4: UVSD $v(old)$ estimates for old/new discrimination. Rearranged pairs were more varied in terms of confidence than intact pairs. Unusually, old WD-Name and BD pairs were associated with lower variance in confidence ratings than pairs of new items. This pattern is difficult to explain in terms of combining different sources of evidence, a common dual-process interpretation of the UVSD model (Wixted, 2007a).

ferences in the variance ratio $v(old)$ from the UVSD model, that is to say we compared the variance of the old pair distribution (relative to the new pair distribution, set for each pair type to a variance of 1) across test condition (*intact*, *rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*)[1]. Repeated measures ANOVA revealed main effects of both test condition [$F(1,23) = 28.46, p < .001$] and pair type [$F(2,46) = 23.59, p < .001$]. Rearranged pairs were associated with greater variance than intact pairs and Bonferroni-corrected paired t-tests revealed the pair type effect was driven mainly by greater variance to WD-Image pairs (both $p < .001$), with only marginally greater WD-Name than BD variance ($p = .083$). An interaction between pair type and test condition [$F(2,46) = 4.06, p = .024$] reflected smaller intact/rearranged differences for WD-Image than BD or WD-Name pairs.

Taken at face value, variance in confidence ratings might reflect correspondingly varied sources of memory evidence. In this context it makes sense that rearranged pairs should lead to greater variance; their components are more independent in terms of memory strength since they were encoded during separate episodes. More surprising is the fact that the variance was greater for new than old items (except for WD-Image pairs). This is the opposite pattern to that normally observed, and is in tension with a common explanation for the UVSD model, i.e. that targets have more varied memory strength as a result of combining multiple sources of evidence (Wixted, 2007a).

The effect of pair type might be related to memory strength overall; the pattern could conceivably arise because hits to studied images were reduced relative to those for studied names, leading to greater variance in their confidence rating on the 10 point scale (but only if correct rejections were not reduced to the same extent). In general, however, effects involving pair type should be interpreted with caution: the ratings used to derive the variance do not only reflect old/new, but also intact/rearranged confidence, which differed across pair type. The results are summarised in Figure 9.4.

[1]Since $v(old)$ is a ratio, all statistical tests and means were calculated using the log-transformed data.

9.3.2 Qualitative intact/rearranged differences across pair types

The associative recognition data was fit to the DPMSD model and the significance of the variance ratio $v(R)$ and familiarity d'_F parameters were tested using likelihood ratios. As in the previous chapters, these analyses supported the inclusion of a variance ratio parameter $v(R)$ for each participant [$\chi^2(24) = 50.00, p = .001$]; recollected trials were associated with a smaller variance in memory strength than non-recollected trials (mean $v(R) = 0.70$). Consistent with the associative recognition data reported previously in this thesis, allowing familiarity to contribute did not significantly improve the fit of the model [$\chi^2(108) = 12.81, p > .999$]. The final model included 9 criteria and 10 memory-related parameters per participant: $p(R_a)$, $p(R_r)$ and d'_R (three values per participant, corresponding to each pair type) and $v(R)$ (one value per participant).

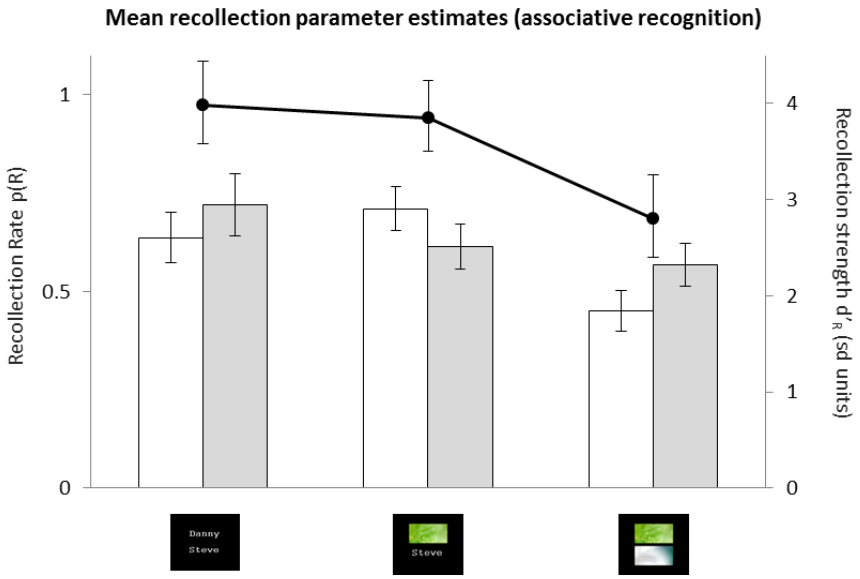


Figure 9.5: Recollection estimates for intact/rearranged discrimination. WD-Name and BD pairs did not differ in terms of overall recollection rate and strength, but recollection rates to intact relative to rearranged BD pairs were significantly greater than they were for WD pairs.

Recollection was, according to the DPMSD model, the sole basis for intact/rearranged discrimination; estimates of its frequency and strength are presented in Figure 9.5. Mean recollection strength differed according to pair type, [$F(2, 46) = 5.41, p = .008$], and according to Bonferroni-corrected paired t-tests this was driven by lower strength to WD-Image than WD-Name ($p = .028$) and, marginally, BD pairs

($p = .087$), but not by differences between these latter two pair types ($p > .999$). We analysed recollection rates using two-way repeated measures ANOVA with factors of test condition (*intact*, *rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*). This revealed a main effect of pair type [$F(1.61, 36.97) = 7.17, p = .004$], reflecting significantly lower recollection rates for WD-image than WD-Name ($p = .011$) or BD ($p = .042$) pairs, which did not differ from each other (Bonferroni-corrected $p > .999$). It also revealed an interaction between pair type and test condition [$F(1.61, 36.92) = 3.90, p = .037$], which reflected the fact that for the BD condition recollection was more frequent for intact than rearranged pairs, whereas for both WD conditions the opposite pattern was observed.

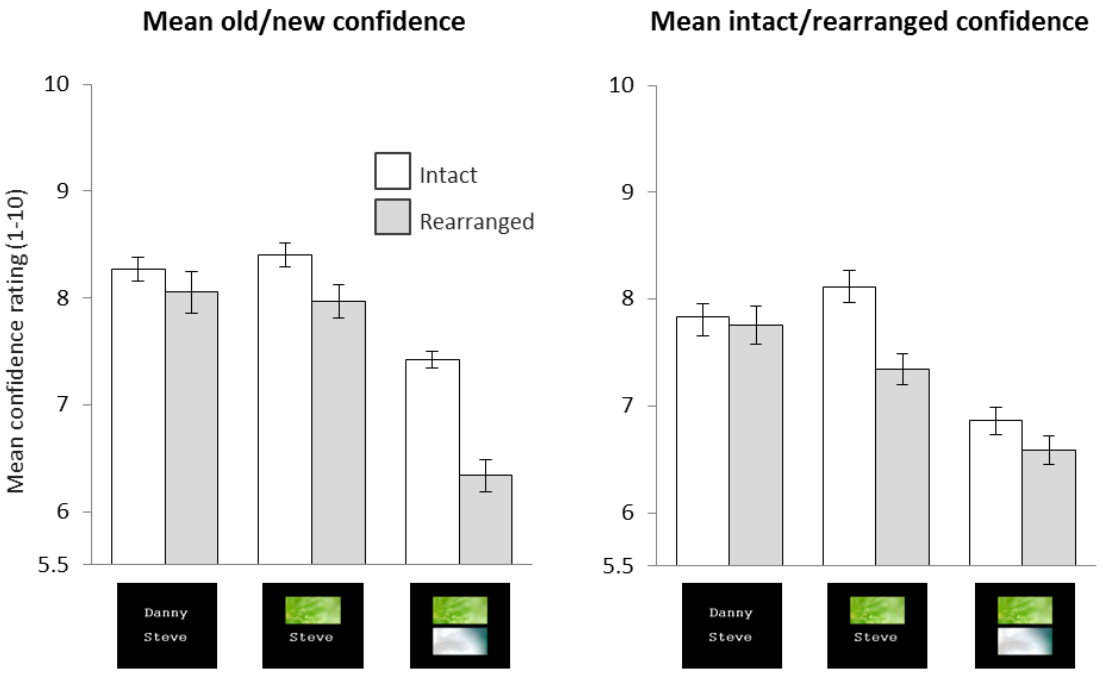
9.3.3 Different patterns of confidence across pair types

The interaction between pair type and test condition found above was not model-specific. We also ran two-way repeated measures ANOVA on both mean old/new and intact/rearranged confidence[2], (data summarised in Figure 9.6). The analysis included factors of test condition (*intact*, *rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*).

The ANOVA on old/new confidence (Figure 9.6(a)) revealed greater confidence to intact than rearranged pairs [$F(1, 23) = 21.21, p < .001$], as well as a main effect of pair type [$F(1.42, 32.58) = 74.48, p < .001$]. Post-hoc Bonferroni-corrected paired t-tests revealed that the main effect of pair type was driven by reduced confidence for WD-Image pairs relative to both WD-Name and BD pairs (both $p < .001$), and that WD-Name and BD pairs were closely matched in terms of overall confidence ($p > .999$). Finally, a strong interaction between test condition and pair type [$F(1.29, 29.76) = 15.65, p < .001$] was revealed by follow-up paired t-tests to reflect greater confidence to intact than rearranged BD pairs ($p = .004$), and the same effect to an even greater extent for WD-Image pairs ($p < .001$), but not WD-Name pairs ($p = .214$).

We repeated the ANOVA above, but this time on the mean confidence to intact/rearranged decisions (Figure 9.6(b)). This revealed main effects of test con-

[2]Participants made a confidence rating from 1-5 after discriminating between intact/rearranged/new. Thus the full confidence scale comprises 10 points, from -5 (confidence 5 for an incorrect decision) up to 5 (confidence 5 for a correct decision), with no zero. For simplicity this is rescaled here to 1-10.

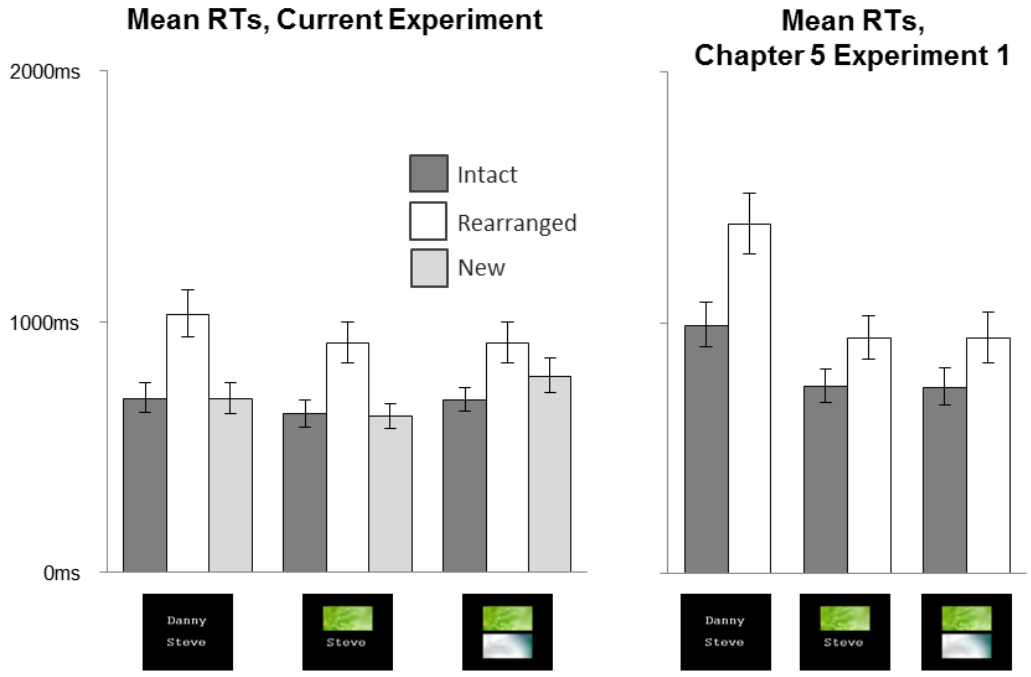


(a) Mean old/new confidence.

(b) Mean intact/rearranged confidence.

Figure 9.6: Mean confidence for old/new and intact/rearranged discrimination. (a) Intact BD or WD-Image pairs were more confidently identified as old than rearranged pairs of the same type, but intact and rearranged WD-Name pairs were identified as old with equal confidence. (b) For WD pairs, intact and rearranged pairs were correctly identified with equivalent confidence. In contrast, intact BD pairs were more confidently identified than rearranged pairs - possibly reflecting greater availability of, or reliance on, holistic characteristics for these pairs. Confidence ratings ranged from 1-10, with a mean rating of 5.5 indicating chance performance.

dition $[F(1,23) = 9.75, p = .005]$ and pair type $[F(2,46) = 47.03, p < .001]$. Intact pairs were associated with higher confidence than rearranged pairs overall, while post-hoc Bonferroni-corrected paired t-tests revealed that the main effect of pair type was driven by reduced confidence for WD-Image pairs relative to both WD-Name and BD pairs (both $p < .001$), whereas WD-Name and BD pairs were again closely matched in terms of overall confidence ($p > .999$). A significant interaction term $[F(2,46) = 6.22, p = .004]$ reflected a greater effect of test condition for BD than WD pairs. Follow-up paired t-tests for each pair type revealed that in fact the test condition effect was isolated to BD pairs ($p < .001$), and not present for WD-Name ($p = .690$) or WD-Image pairs ($p = .127$). This latter interaction can be considered analogous to the conclusion from the DPMSD model above, namely



(a) Mean RTs, current experiment.

(b) Mean RTs, Chapter 5, Experiment 1.

Figure 9.7: Mean reaction times for both ERP experiments. (a) In the current experiment, correct responses to rearranged pairs were slower than to either intact or new pairs, with slightly longer response times to WD-Name pairs than other pair types. (b) Differences between intact and rearranged response times were still present when no new items were shown at test, and WD-Name pairs were again identified more slowly than other pair types. Note that response times are relative to the onset of the response screen, not the test presentation.

that BD (but not WD) pairs showed greater recall-to-accept than recall-to-reject rates.

9.3.4 Reaction time differences across pair types

Reaction times are particularly relevant in ERP studies, since ERP memory effects are normally characterized at least partly on the basis of their time course. Thus, the interpretation of ERP differences across conditions depends on the existence or otherwise of RT differences: these may reflect either the engagement of different cognitive processes, or changes to the time course of the same processes.

We therefore analysed reaction times for correct responses in both experiments[3], summarised in Figure 9.7. We used repeated measures ANOVA separately for each experiment with factors of test condition (*intact*, *rearranged*, *new*) and pair type (*WD-Name*, *WD-Image*, *BD*). For the current experiment, Figure 9.7(a), this revealed a main effect of test condition [$F(2,46) = 30.68, p < .001$], driven by increased reaction times to rearranged compared to intact and new pairs (both $p < .001$], which did not significantly differ ($p > .999$]. There was a marginal main effect of pair type [$F(2,46) = 3.10, p = .055$]; Bonferroni-corrected paired t-tests suggested this reflected longer reaction times to WD-Name than BD pairs ($p = .036$] but no differences involving WD-Image pairs (both $p > .203$]. Finally, there was a significant interaction between pair type and test condition [$F(1.60, 41.70) = 42.70, p < .001$], which reflected longer reaction times to new WD-Image pairs.

We also considered reaction times from Chapter 5 (Experiment 1), as shown in Figure 9.7(b). Repeated measures ANOVA with factors of test condition (*intact*, *rearranged*) and pair type (*WD-Name*, *WD-Image*, *BD*) revealed main effects of both: intact pairs were recognised more quickly than rearranged pairs [$F(1,26) = 50.69, p < .001$] and responses were slowest to WD-Name pairs [$F(1.60, 41.70) = 42.70, p < .001$]. Bonferroni-corrected paired t-tests confirmed that WD-Name pairs were associated with slower responses than WD-Image or BD pairs (both $p < .001$] but that WD-Image and BD pairs did not significantly differ from each other ($p > .999$). Reaction times were clearly somewhat slower in Chapter 6 (Experiment 1) than the current study, most likely reflecting an earlier response screen onset (500ms and 1000ms after test presentation onset, respectively).

In summary, intact pairs were generally recognised as old more confidently than rearranged pairs were, varying apparently with item type: WD-Image, and to a smaller extent BD pairs, showed greater old/new confidence differences between intact and rearranged pairs (Figure 9.6(a)). Reaction times were fastest for intact and new pairs; responses to rearranged pairs were on average 250ms slower. This may suggest that additional or longer cognitive processes were invoked when rearranged pairs were recognised, but they also mean that ERP differences between rearranged and new/intact pairs should be carefully examined in case they are

[3]The reaction time data was log-normally distributed; we therefore report geometric means and all statistical analyses are performed on the log-transformed data.

driven mainly by differences in the time course of processing, rather than differential processing per se. Mean intact/rearranged confidence was also greater for intact than rearranged pairs, but only for BD pairs: an interaction most dramatically illustrated by the difference between BD and WD-Name pairs, despite near-identical overall performance, mean confidence and reaction times (Figure 9.6(b)). According to the DPMSD model, the reason for this interaction in confidence strength was more frequent recollection to intact than rearranged BD pairs, but the opposite pattern for WD pairs (Figure 9.5). If this is the case, intact/rearranged ERP differences which are larger for BD than WD pairs might be related to recollection success. Accurate quantitative estimates of familiarity were impossible to extract from the data, but there was evidence that familiarity contributed to old/new decisions, given that the DPMSD estimates suggested virtually every trial contained some information which was diagnostic of previous occurrence. In contrast, information which discriminated between intact and rearranged pairs was only present on around 60% of trials. If this interpretation of the DPMSD model is accurate, we should expect to see evidence of the FN400 in contrasts between old and new items, but not in contrasts between intact and rearranged pairs.

9.4 Electrophysiological results

Grand average ERPs for the current experiment were constructed for correct responses to intact, rearranged and new pairs, and are shown separately for each condition in Figures 9.8–9.10. We first conducted a broad analysis to characterise old/new effects across time, test condition, pair type and scalp position. To this end, we conducted six ANOVAs, separately for intact and rearranged pairs and for each of three epochs (300-500ms, 500-800ms and 800-1100ms).

Each ANOVA included factors of test condition (*old (intact or rearranged)*, *new*), pair type (*WD-Names*, *BD*, *WD-Images*), location (*frontal*, *fronto-central*, *central*, *centro-parietal*, *parietal*), hemisphere (*left*,*right*) and site (*superior*, *mid*, *inferior*). To be clear, the test condition (old/new) factor is referred to throughout as either intact/new or rearranged/new, as appropriate. Interactions are reported where they involve the effect of interest (i.e. the old/new effect), and are interpreted on visual examination of the data.

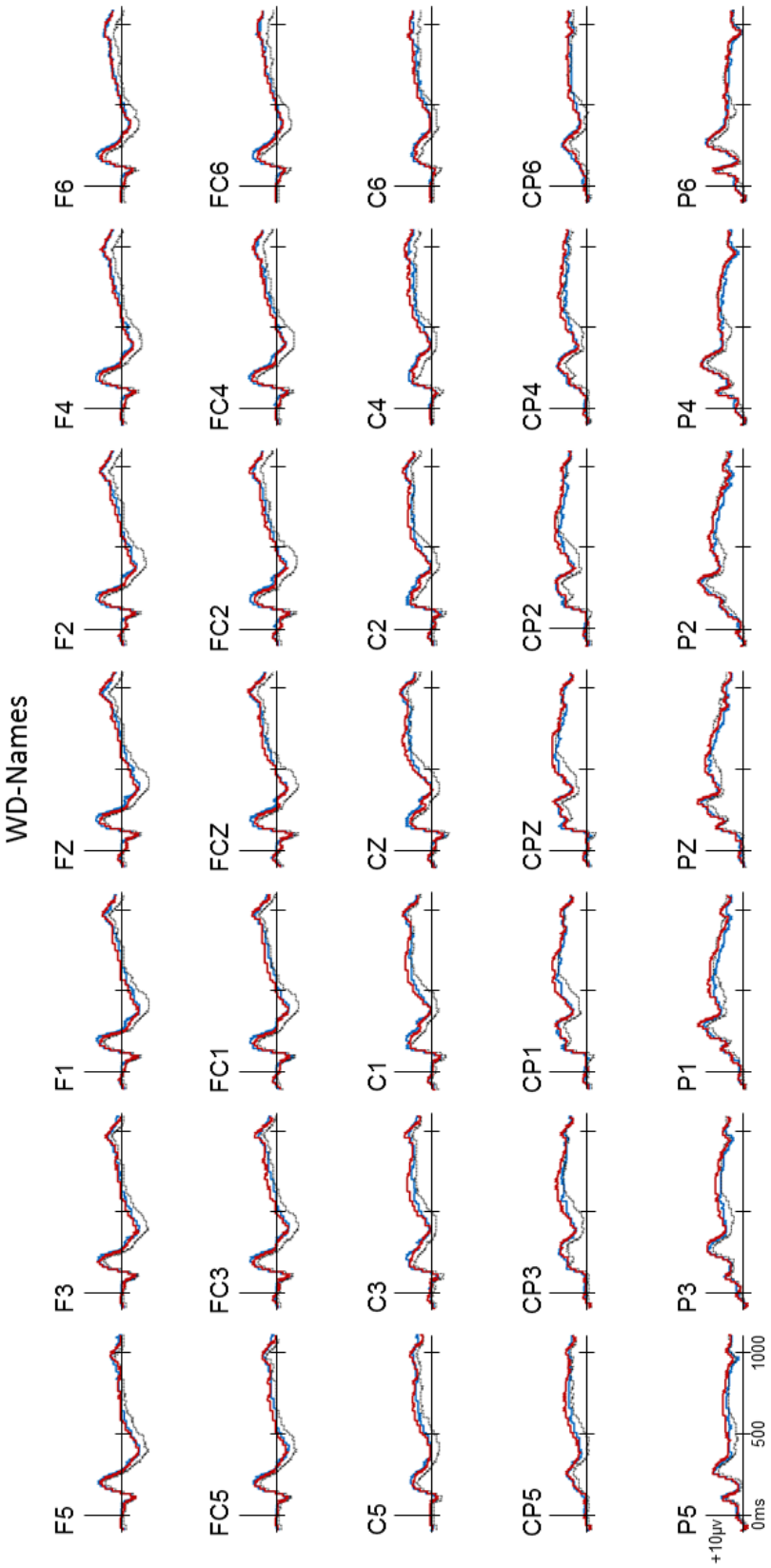


Figure 9.8: Grand average intact (red), rearranged (blue) and new (black) ERPs for WD-Name pairs.

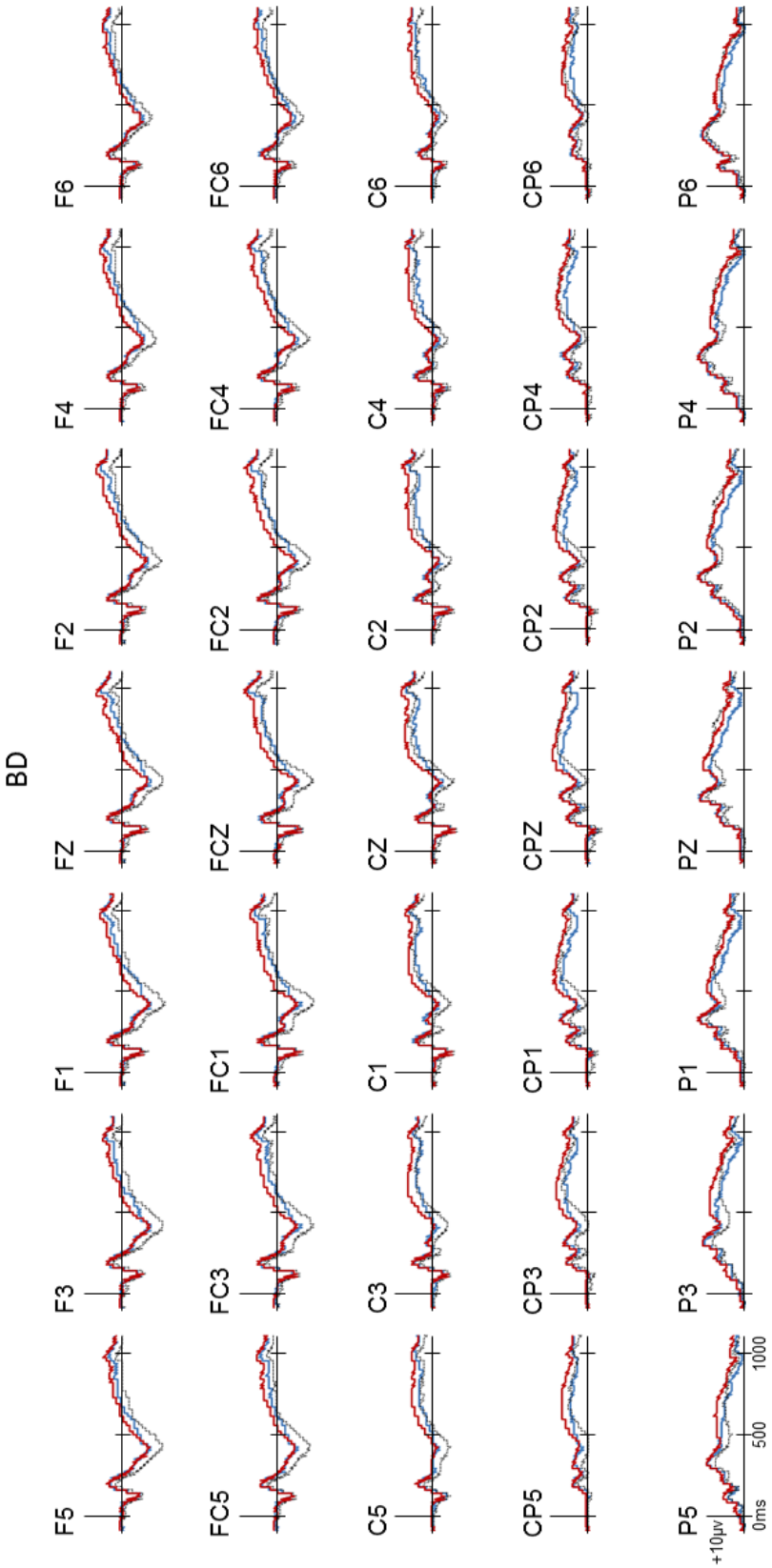


Figure 9.9: Grand average intact (red), rearranged (blue) and new (black) ERPs for BD pairs.

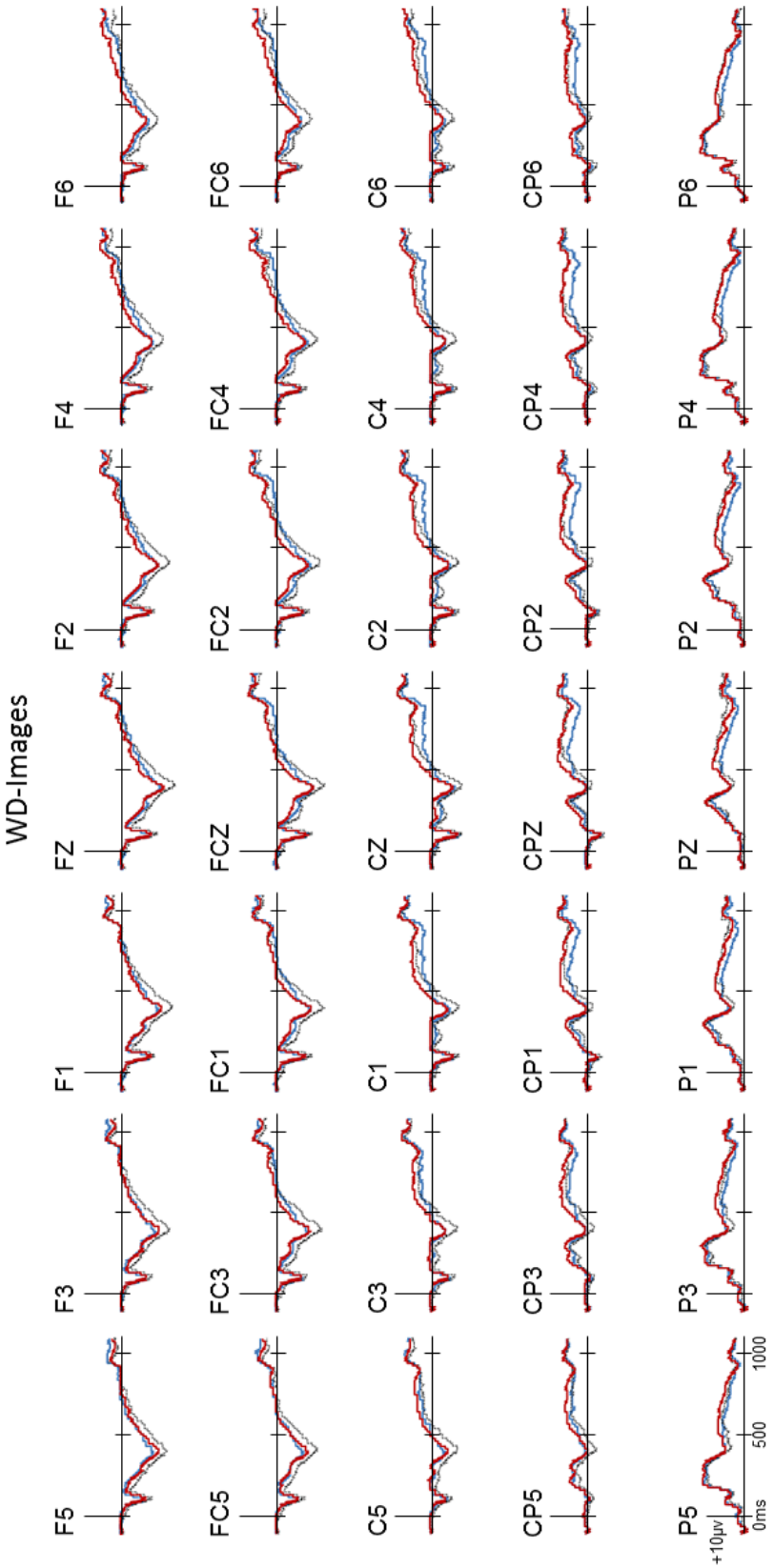


Figure 9.10: Grand average intact (red), rearranged (blue) and new (black) ERPs for WD-Image pairs.

9.4.1 Early (300-500ms) old/new effects

The intact/new and rearranged/new effects in this time window are shown topographically in Figure 9.11. Old pairs of both kinds elicited consistently more positive-going scalp potentials relative to pairs of new items. This positivity had an anterior focus, and thus appears to be broadly consistent in polarity, time course and topography with the FN400 effect linked to familiarity. A first ANOVA, comparing intact and new pairs between 300-500ms, revealed a significant main effect of intact/new $[F(1,23) = 41.98, p < .001]$, indicating that ERPs to intact pairs were more positive going than those to new pairs. The intact/new effect interacted with location $[F(1.26, 28.98) = 9.65, p = .002]$ and site $[F(1.12, 25.76) = 8.57, p = .006]$; the greatest intact/new difference was found frontally and centrally, and was approximately symmetrical across hemispheres. Pair type did not interact significantly with the intact/new effect.

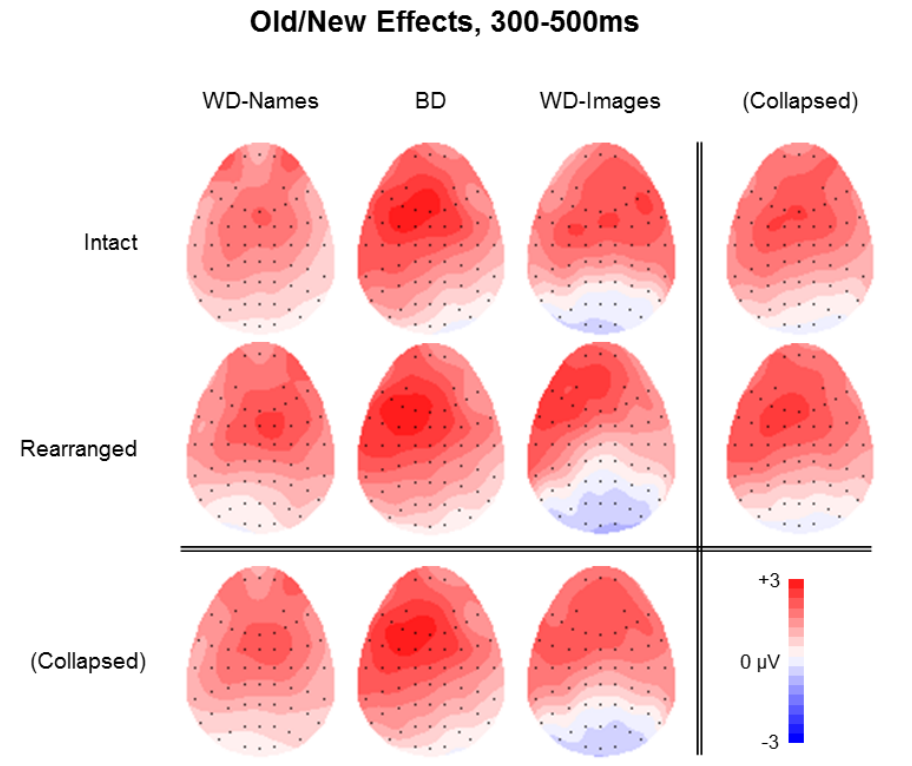


Figure 9.11: Topographic distribution of old/new effects between 300-500ms. The old/new effects are shown separately for test condition of old pairs (*intact*, *rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*), and collapsed across each factor (bottom row and right column respectively).

The comparison between rearranged and new pairs produced similar, but slightly more complex, results. Just as for the intact/new comparison above, rearranged pairs gave rise to significantly more positive going ERPs than new pairs [$F(1,23) = 26.01, p < .001$] and a significant interaction with location [$F(1.172, 26.96) = 7.673, p = .008$] reflected a frontal distribution for the effect. Unlike for intact pairs, however, the rearranged/new effect did not interact with site, indicating that it was comparatively less focused towards the midline. More importantly, analysis also revealed interactions involving pair type. A three-way interaction between rearranged/new effect, pair type and hemisphere [$F(1.57, 36.20) = 4.17, p = .032$] mainly reflected a slight right-hemisphere bias for WD-name pairs, as compared to a more left-of-midline effect for BD and WD-Image pairs. A three-way interaction between rearranged/new, hemisphere and site [$F(1.24, 28.57) = 4.31, p = .039$] reflected an overall bias of the rearranged/new effect toward the left of the scalp, in that left hemisphere effects were distributed somewhat towards inferior sites, while right hemisphere effects were biased towards superior sites. This further interacted with pair type [$F(2.50, 57.45) = 3.94, p = .018$], suggesting some differences in this pattern between WD-Name and BD pairs (which were relatively superior) and WD-Image pairs (which had a more inferior focus, but on the same - left - hemisphere as BD pairs).

To better quantify this pattern and to reduce the chances of accepting artifactual topographic differences driven by differences in mean effect size, we conducted a topographic analysis of the distribution of the early rearranged/new effect across conditions. We subtracted the new from rearranged waveforms, and then rescaled the resulting differences (Max-Min method, see Section 8.3.4). The resulting ANOVA included factors of pair type, location, hemisphere and site, all defined as for the initial analysis above, and confirmed the effects found above involving pair type: an interaction with hemisphere [$F(1.66, 38.19) = 4.29, p = .027$], and an interaction with both hemisphere and site [$F(2.56, 58.95) = 3.82, p = .019$].

In summary, the initial analysis confirmed the presence of characteristic frontal positivity associated with familiarity in the 300-500ms time window, i.e. the FN400. As can be seen in Figure 9.11 the effect was frontally distributed for both rearranged and intact pairs, and for all three pair types. All three pair types produced consistent intact/new effects, which were focused on superior sites, but there was evidence of differences across conditions for rearranged pairs. Here the

effects were less centrally focused, and the exact lateral distribution of the effect also appeared to vary by condition: WD-Name pairs gave rise to an effect with a slight right bias, WD-Image pairs produced a markedly left-biased effect and BD pairs lay somewhere in between, peaking at superior left hemisphere electrodes.

9.4.2 Mid (500-800ms) old/new effects

The intact/new and rearranged/new effects in this time window are shown topographically in Figure 9.12. These show a greater positivity to old than new pairs at the front of the scalp, while at the back there is a posterior negativity (greater for rearranged than intact pairs). This pattern, however, seems to be accompanied by an left-sided positivity to old pairs at parietal electrodes, possibly reflecting the onset of the LPONE. A first ANOVA comparing intact and new pairs between 500-800ms revealed a significant main effect of intact/new [$F(1,23) = 15.53, p = .001$]; ERPs to intact pairs were more positive going than those to new pairs. The intact/new effect interacted with location [$F(1.47,33.76) = 9.49, p = .001$] and hemisphere [$F(1,23) = 10.56, p = .004$]; the effect was most strongly positive at the front and left of the scalp, while right occipital regions showed a smaller or slightly negative intact/new deflection. A three-way interaction between intact/new, location and hemisphere [$F(1.41,32.50) = 7.23, p = .006$] indicated that the left-hemisphere bias was driven by greater positivity at left parietal locations, and was not present at the front of the scalp. Another three-way interaction, between intact/new, hemisphere and site [$F(1.14,26.11) = 8.25, p = .006$] may have reflected an overall bias of the effect to the left of the scalp; differences at the left hemisphere were focused more strongly on inferior sites than those at the right hemisphere. We also found two significant interactions involving pair type: between intact/new, hemisphere, site and pair type [$F(2.38,54.69) = 4.12, p = .016$] and between intact/new, hemisphere, location, site and pair type [$F(7.19,165.28) = 2.54, p = .016$].

To characterise the interactions between intact/new and pair type we rescaled the data and conducted a topographic analysis in identical fashion to the previous section. This analysis confirmed the two interactions: Hemisphere by site by pair type [$F(2.11,48.47) = 3.19, p = .048$] and location by hemisphere by site by pair type [$F(7.20,165.57) = 2.34, p = .025$]. The first reflected a different distribution

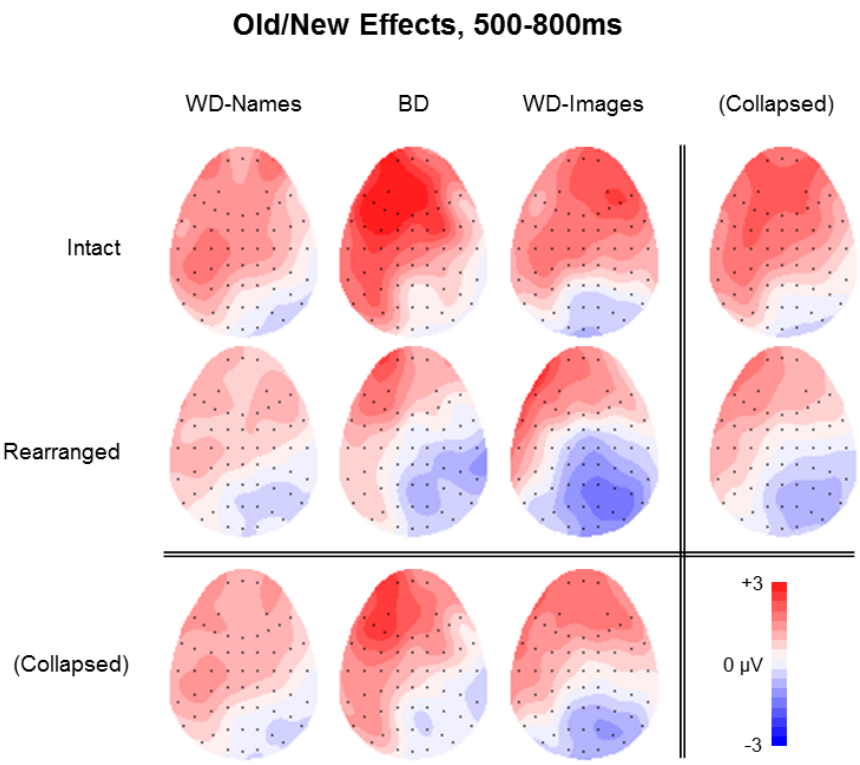


Figure 9.12: Topographic distribution of old/new effects between 500-800ms. The old/new effects are shown separately for test condition of old pairs (*intact*, *rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*), and collapsed across each factor (bottom row and right column respectively).

on the right half of the scalp for WD-Image pairs (for which the intact/new differences were evenly distributed across superior and inferior sites) and the other two pair types (for which the differences were focused on superior sites). The second, four-way interaction is harder to interpret, though it appears to reflect that pair-type specific differences in lateral distributions were clearest at frontal electrodes (here BD and WD-Image pairs differ in hemispheric focus, while WD-Name pairs show a more central, i.e. superior focus than either). At posterior electrodes, in contrast, the lateral distribution of intact/new differences were more consistent across conditions.

We next analysed rearranged/new effects in this time window using five-way ANOVA as above. Unlike for intact pairs, this revealed no main effect of rearranged/new, indicating that any positive or negative deflections were relatively localised and counterbalanced each other. This was confirmed by significant in-

teractions of rearranged/new with location [$F(1.21, 27.85) = 8.53, p = .005$] and hemisphere [$F(1, 23) = 21.90, p < .001$], which reflected the fact that the rearranged/new difference was positive going at the front and left of the scalp respectively, but generally negative-going at right-parietal electrodes. Finally, a three-way interaction between rearranged/new, hemisphere and site [$F(1.21, 27.74) = 20.29, p < .001$] reflected an overall left bias for the effect, in that inferior sites were more positive than superior sites on the left of the scalp, but vice-versa on the right. Pair type did not interact significantly with the rearranged/new effect.

To summarise, from 500-800ms intact pairs produced more positive going ERPs than new pairs, with the greatest positivity located at frontal and left hemisphere electrodes. The distribution of this intact/new effect varied according to pair type: WD-Name pairs showed a left-parietal focus, while WD-Image and BD pairs also showed (different) patterns of positivity at the front of the scalp, and smaller or slightly negative effects at superior posterior regions. Nonetheless, all three pair types showed some evidence of a left-parietal old/new effect, indexed by significantly greater intact/new positivity at left than right parietal electrodes.

Rearranged pairs showed a similar pattern, but with lower overall positivity, such that ERPs to rearranged pairs were no more positive on average (across the scalp) than new pairs. Frontal and left hemisphere electrodes showed greater positivity than right-parietal electrodes, which were consistently below zero. Unlike for intact pairs, however, positivity was greatest on the left of the scalp at both anterior and posterior locations, and this did not differ across pair type. While this constituted some evidence of a left-parietal rearranged/new effect, it was also accompanied by positivity at left frontal electrodes.

9.4.3 Late (800-1100ms) old/new effects

We also examined effects later in the epoch, (illustrated in Figure 9.13), which may be informative in this experiment given that the task performed by participants is a relatively complex and potentially two-stage decision (old v new; intact v rearranged). During this 800-1100ms time window, both rearranged and, to a lesser extent, intact pairs showed a superior-forcused negative posterior deflection relative to new pairs. This appeared to be accompanied by more positive ERPs to old than new pairs at the front of the scalp, with relatively unfocused lateral posi-

tion. Firstly, ANOVA comparing intact and new pairs revealed no main effect of intact/new, but intact/new did interact with location [$F(1.37, 31.46) = 17.15, p <$ .001] and site [$F(1.19, 27.46) = 6.37, p = .014$], reflecting a negative deflection for intact pairs at posterior and superior locations. A significant interaction between intact/new, location and site [$F(2.58, 59.32) = 8.00, p < .001$] emphasised that this deflection was driven specifically by superior parietal electrodes - it did not extend to superior frontal or inferior parietal electrodes. An interaction between intact/new, location, hemisphere and site [$F(2.88, 66.22) = 4.49, p = .007$] reflected positivity at inferior right frontal electrodes, but not at corresponding electrodes on the left hemisphere. Two complex interactions involving pair type hinted at possible differences in the distribution of effects across BD and WD pairs. Firstly, an interaction between intact/new, pair type, hemisphere and site [$F(2.17, 50.00) = 3.56, p = .032$] appears to reflect a right-inferior, but left-superior focus (i.e. a general right-sided bias) for the intact/new effect in both WD conditions, while the effect is more symmetrical for BD pairs. This further interacted with location [$F(7.76, 178.37) = 3.17, p = .002$], possibly suggesting that these differences were driven primarily by the variability seen in frontal electrodes in Figure 9.13.

We checked the interactions involving pair type using a topographic analysis of the rescaled intact/new effect across conditions. The same interactions involving pair type were present: with hemisphere and site [$F(2.14, 49.22) = 3.28, p = .043$], reflecting differences in lateral distribution of the intact/new effect across pair type, and with hemisphere, site and location [$F(7.85, 180.63) = 3.15, p = .002$], likely reflecting a greater positivity for BD than WD pairs at left-frontal, but not parietal or right-frontal electrodes (Figure 9.14). The presence of site in the interaction terms indicates that the lateral differences between conditions are not symmetrical around the midline; WD-Image pairs for example peak at more inferior sites on the right of the scalp than BD pairs do on the left. This explanation, i.e. that WD and BD pair types differed in lateral distribution at frontal but not parietal locations, is also consistent with two marginal interactions: that of pair type with location and hemisphere [$F(3.14, 72.13) = 2.59, p = .057$] and of pair type with location and site [$F(3.83, 88.02) = 2.32, p = .066$].

Analysis of the rearranged/new effect in this epoch revealed a similar broad pattern as for intact/new: strong interactions between rearranged/new and loca-

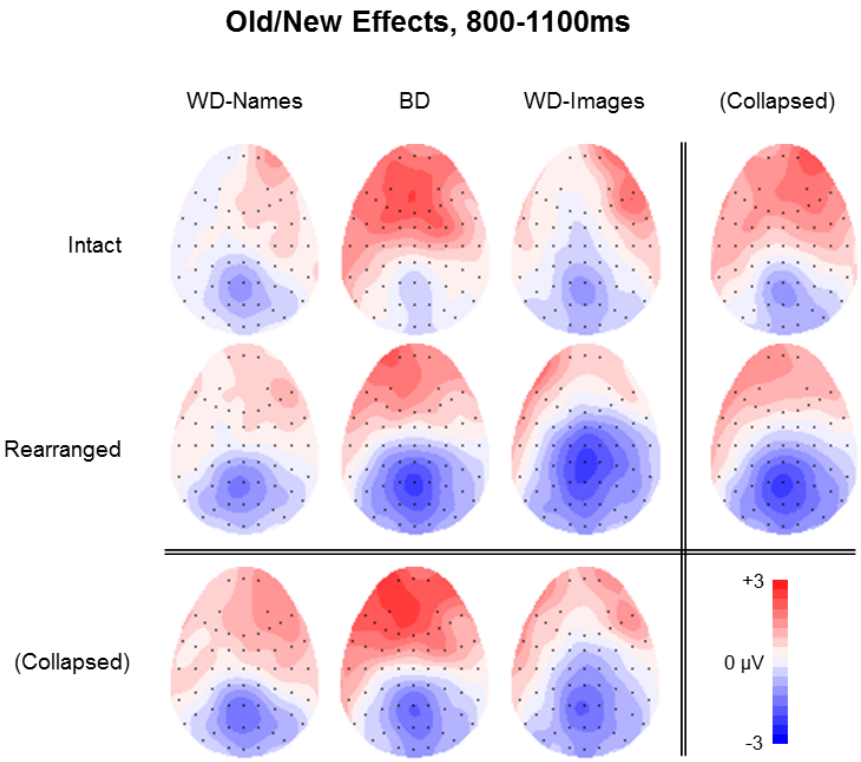


Figure 9.13: Topographic distribution of old/new effects between 800-1100ms. The old/new effects are shown separately for test condition of old pairs (*intact*, *rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*), and collapsed across each factor (bottom row and right column respectively).

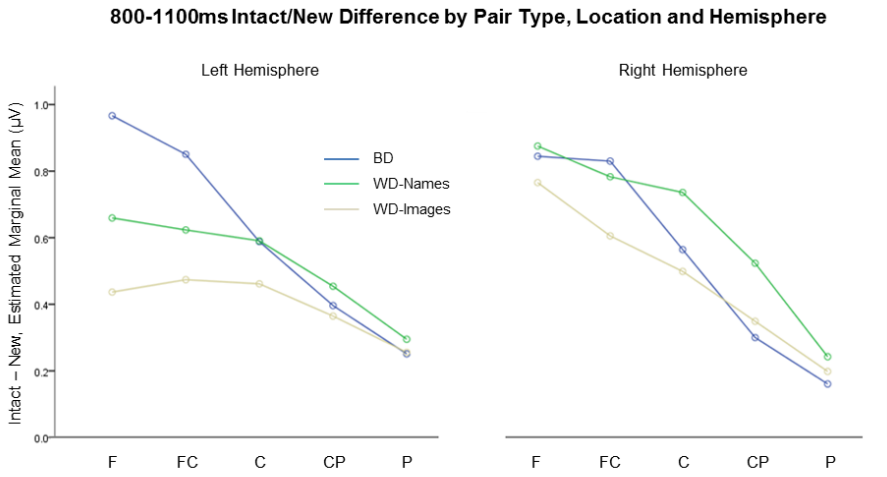


Figure 9.14: Increased late left-frontal positivity to intact BD pairs. Estimated marginal means for each pair type show matched intact/new differences on the right of the scalp, but greater positivity for BD than WD pairs on the left, between 800-1100ms.

tion [$F(1.28, 29.46) = 19.09, p < .001$], rearranged/new and site [$F(1.08, 24.91) = 17.62, p < .001$], and between all three of these factors [$F(3.41, 78.48) = 12.82, p < .001$]. Together this reflected a superior parietal negativity and inferior frontal positivity elicited by rearranged pairs relative to new pairs, which was broadly symmetrical around the midline. Unlike for intact pairs, however, the rearranged/new effect did not vary with pair type. Finally, there was also an interaction between rearranged/new, hemisphere and site [$F(1.42, 32.75) = 6.07, p = .011$] which appeared to reflect an inferior positivity to rearranged pairs on the left hemisphere, which was less strong or entirely absent on the right hemisphere.

Analyses between 800-1100ms thus confirm what is clearly visible in Figure 9.13: old pairs were associated with a relatively focused superior parietal negativity relative to new pairs. It is possible that effects in later epochs which are symmetrical around the midline, such as the strong negativity we observed here, may reflect differences in evoked potentials associated with motor activity or response preparation (Kuo and Van Petten, 2006). This can occur when the timing of these potentials differs across two conditions, creating large amplitude differences in the contrast which occur at superior sites if response hands for each condition are counterbalanced across participants. This can be tested by comparing ERPs to participants responding with one hand/condition mapping versus those with

the opposite; if the ERP differences are related to motor preparation they should be lateralized and on opposite hemispheres for the two groups. We divided participants into two groups according to the hand used for each type of response, and included it as a between-subjects factor in a repeated measures ANOVA for old/new differences between 800-1100ms. The ANOVA also included factors of hemisphere[4] and pair type, and the analysis was run separately for intact and rearranged pairs. The crucial hemisphere by group interaction was not significant for either intact ($p = .851$) or rearranged pairs ($p = .738$), meaning that the old/new effect was not driven by differences in lateralized readiness potentials.

In summary, both intact and rearranged pairs showed focused negativity compared to new pairs at superior parietal electrodes, consistent in polarity, distribution and timing with the LPN effect, as well as a more laterally distributed positivity at frontal electrodes (Figure 9.13). For intact pairs, the lateral distribution of this frontal positivity differed across pair types, with both WD-Name and WD-Image pairs showing peak intact/new effects at right frontal sites, while BD pairs showed positivity at mid/left frontal electrodes. Estimated marginal means for each pair type suggested that this difference was driven primarily by increased front-left activity to BD but not WD pairs (see Figure 9.14). By contrast, rearranged pairs showed widely distributed frontal positivity compared to new pairs, with a somewhat left inferior focus, which did not differ by pair type.

9.4.4 Summary of old/new ERP effects

The analyses above were used to characterize ERP differences between pairs of old and new items. They reveal evidence of three old/new ERP effects we introduced in this chapter: the FN400, the LPONE and the LPN. In this section we summarise the magnitude of each of these effects by focusing on old/new differences at the electrode locations and time windows associated with each one.

Firstly, the FN400 is characterised by a positive deflection to old items which is maximal at midfrontal electrodes (here F1, FZ, F2) during 300-500ms (Figure 9.15). A 3×3 repeated measures ANOVA confirmed more positive-going ERPs to

[4]We averaged the activity from 6 electrodes for each hemisphere, C1, C3, CP1, CP3, P1 and P3, in order to capture the relatively parietally located old/new effect.

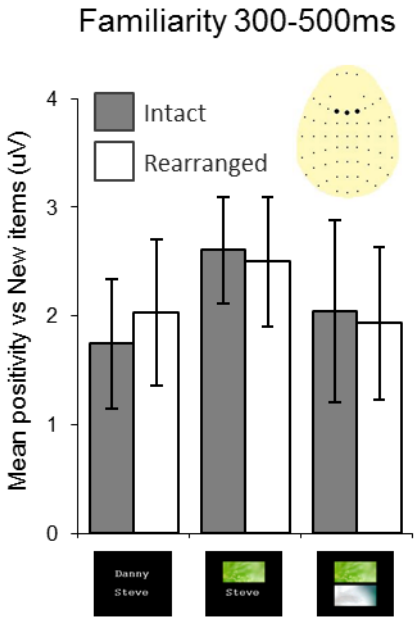
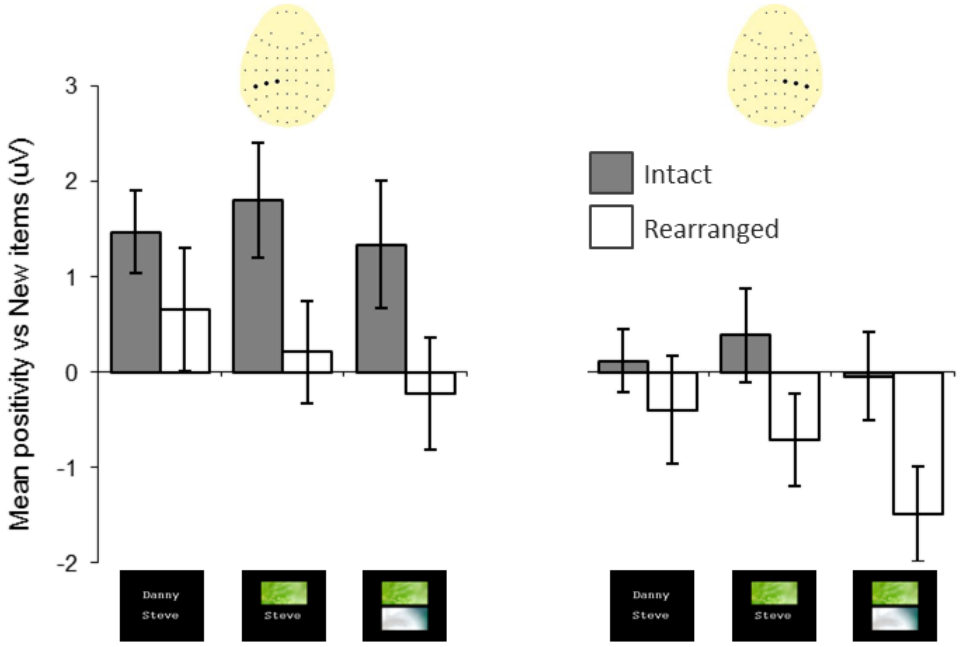


Figure 9.15: Magnitude of the FN400 by test condition and pair type. Magnitudes were calculated by averaging activity across electrodes F1, FZ and F2 between 300-500ms post-stimulus onset. The FN400 did not vary with either intact/rearranged or pair type.

old than to new pairs: a main effect of test condition [$F(2,46) = 20.84, p < .001$] reflected more negative ERPs to new than to intact or rearranged pairs (both $p < .001$), but no difference between intact and rearranged pairs ($p = .945$). Thus, we observed an effect consistent with the FN400: a positivity to old relative to new pairs at midfrontal electrodes between 300-500ms. We also conducted a direct analysis of the old/new differences shown in Figure 9.15 using repeated measures ANOVA with factors of test condition (*intact*,*rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*). Neither factor approached significance and they did not interact (all $p > .678$), indicating that the FN400 was not significantly modulated by intact/rearranged or pair type.

The LPONE is similarly characterised by old/new positivity, but at left-parietal electrodes (here P5, P3, P1) between 500-800ms post-stimulus onset. We conducted a 3×3 repeated measures ANOVA on average activity at these electrodes and during this time window, which revealed a main effect of test condition [$F(2,46) = 8.59, p = .001$] driven by increased positivity to intact than either rearranged ($p = .003$) or new ($p = .001$) pairs, which did not differ from eachother ($p = .596$). Does this mean intact pairs elicited the LPONE, but rearranged pairs



(a) Mean old/new differences, left hemisphere.

(b) Mean old/new differences, right hemisphere.

Figure 9.16: Mean old/new differences at parietal electrodes between 500-800ms. Intact pairs elicited more positive-going ERP effects than rearranged pairs. Importantly, however, old/new differences were greater on the left, (a), than the right, (b), of the scalp but intact/rearranged differences were statistically matched across hemispheres; this pattern suggests that old/new differences were related to the LPONE but intact/rearranged differences were not.

did not? To determine the answer to this question we analysed the old/new differences shown in Figure 9.16 using repeated measures ANOVA; i.e. we included factors of test condition (*intact*,*rearranged*) and pair type (*WD-Names*, *BD*, *WD-Images*), but also hemisphere. This confirmed the two main effects evident in Figure 9.16: intact pairs elicited more positive-going ERPs than rearranged pairs [$F(1,23) = 11.00, p = .003$] and left-hemisphere electrodes, Figure 9.16(a), elicited much greater old/new effects than those on the right, Figure 9.16(b), [$F(1,23) = 51.75, p < .001$]. Crucially, however, intact/rearranged and hemisphere did not interact ($p = .312$), inconsistent with the hypothesis that intact pairs elicited a greater LPONE than rearranged pairs; rather, the difference was distributed evenly across both hemispheres. Thus a more parsimonious explanation for the pattern is that both intact and rearranged pairs showed similar

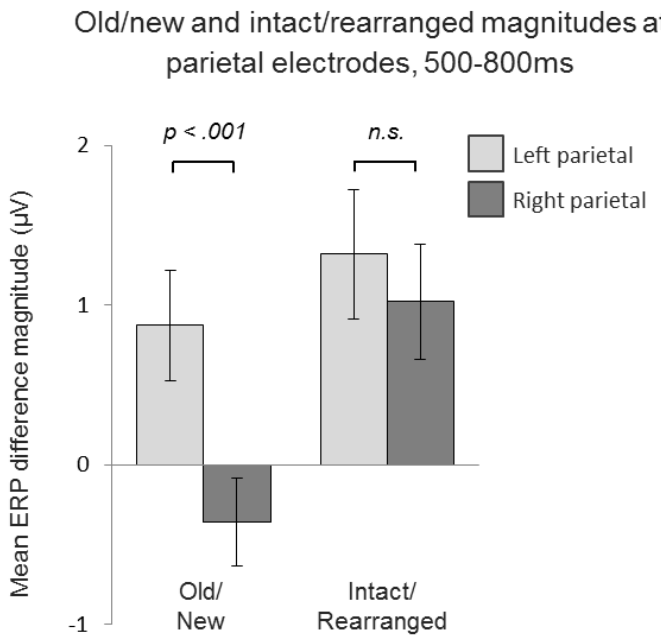


Figure 9.17: Parietal old/new and intact/rearranged differences by hemisphere, between 500-800ms post stimulus onset. Old/new differences are restricted to left-parietal electrodes (here P5, P3, P1; right-parietal = P2, P4, P6) but intact/rearranged effects are symmetrical about the midline, suggesting the LPONE exists in old/new contrasts but does not significantly differ by intact/rearranged.

LPONE effects but rearranged pairs elicited stronger (or earlier-onsetting) LPN effects than intact pairs (see also Figure 9.12). Figure 9.17 shows the pattern clearly, by collapsing ERPs across pair types. Mean magnitudes of both old/new and intact/rearranged differences are shown for left parietal (P5, P3, P1) and right parietal electrodes (P2, P4, P6) between 500-800ms. These demonstrate that the old/new effect is present only on the left hemisphere (consistent with the presence of the LPONE in old/new contrasts) but intact/rearranged differences are symmetrical across the back of the scalp (inconsistent with these differences being related to the LPONE).

LPN effects are normally maximal shortly after the LPONE (800-1100ms) and at superior parietal electrodes (here P1, PZ, P2). We examined LPN magnitudes across test conditions and pair types using these criteria (Figure 9.18). First, 3×3 repeated measures ANOVA on average activity at these electrodes between 500-800ms revealed a main effect of test condition $[F(2,46) = 12.89, p < .001]$ driven by differences across all three conditions: new pairs elicited more positive

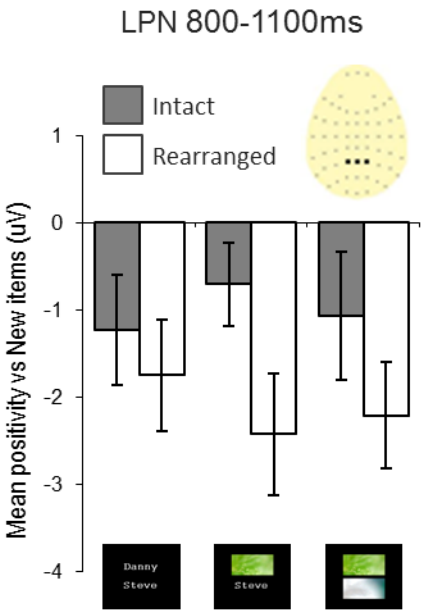


Figure 9.18: Magnitude of the LPN by test condition and pair type. Magnitudes were calculated by averaging activity across electrodes P1, PZ and P2 between 500-800ms post-stimulus onset. LPN magnitudes were greatest for rearranged pairs, but did not differ with pair type.

ERPs than intact pairs ($p = .021$), which were in turn more positive than those for rearranged pairs ($p = .007$; new v rearranged $p < .001$). Repeated measures ANOVA on the old/new difference magnitudes shown in Figure 9.18 confirmed greater rearranged/new than intact/new differences [$F(1,23) = 8.64, p = .007$], but no effects involving pair type: rearranged pairs elicited larger LPNs than intact pairs did.

9.4.5 Early (300-500ms) intact/rearranged effects

So far we have examined old/new differences, enabling us to related our results to well-studied ERP effects. Nevertheless, the primary aim of this chapter is to gain insight into the cognitive processes underlying associative recognition, i.e. the discrimination of intact from rearranged pairs. In the following sections, therefore, we directly compare ERPs to correctly-identified intact and rearranged pairs to identify what significant differences exist between the two conditions, when they first occur and where on the scalp they are located. In doing so we analyse both the current dataset and ERPs to intact/rearranged responses from Chapter 5 (Experiment 1), which are shown in Figures 9.19–9.21.

Early intact/rearranged effects across conditions are summarised in Figure 9.22. We first submitted average electrode potentials from the current experiment to a five-way repeated measures ANOVA, with factors of test condition (*intact*, *rearranged*), pair type (*WD-Names*, *BD*, *WD-Images*), location (*frontal*, *fronto-central*, *central*, *centro-parietal*, *parietal*), hemisphere (*left*,*right*) and site (*superior*, *mid*, *inferior*). This revealed no main effect of intact/rearranged, but two interactions: intact/rearranged, location and hemisphere [$F(1.75, 40.24) = 4.60, p = .020$], and intact/rearranged, pair type, hemisphere and site [$F(1.42, 32.75) = 6.07, p = .011$], though this latter interaction was absent once intact/rearranged effects were rescaled. The former interaction is somewhat difficult to interpret given that marginal means at each location/hemisphere combination showed only very small intact/rearranged differences ($< 0.25\mu V$). We therefore separately analysed the electrode string at each location; none revealed a significant interaction between intact/rearranged and hemisphere.

We also note that the interaction is not present in the equivalent intact/rearranged contrast from Chapter 5 (Experiment 1). An identically structured ANOVA

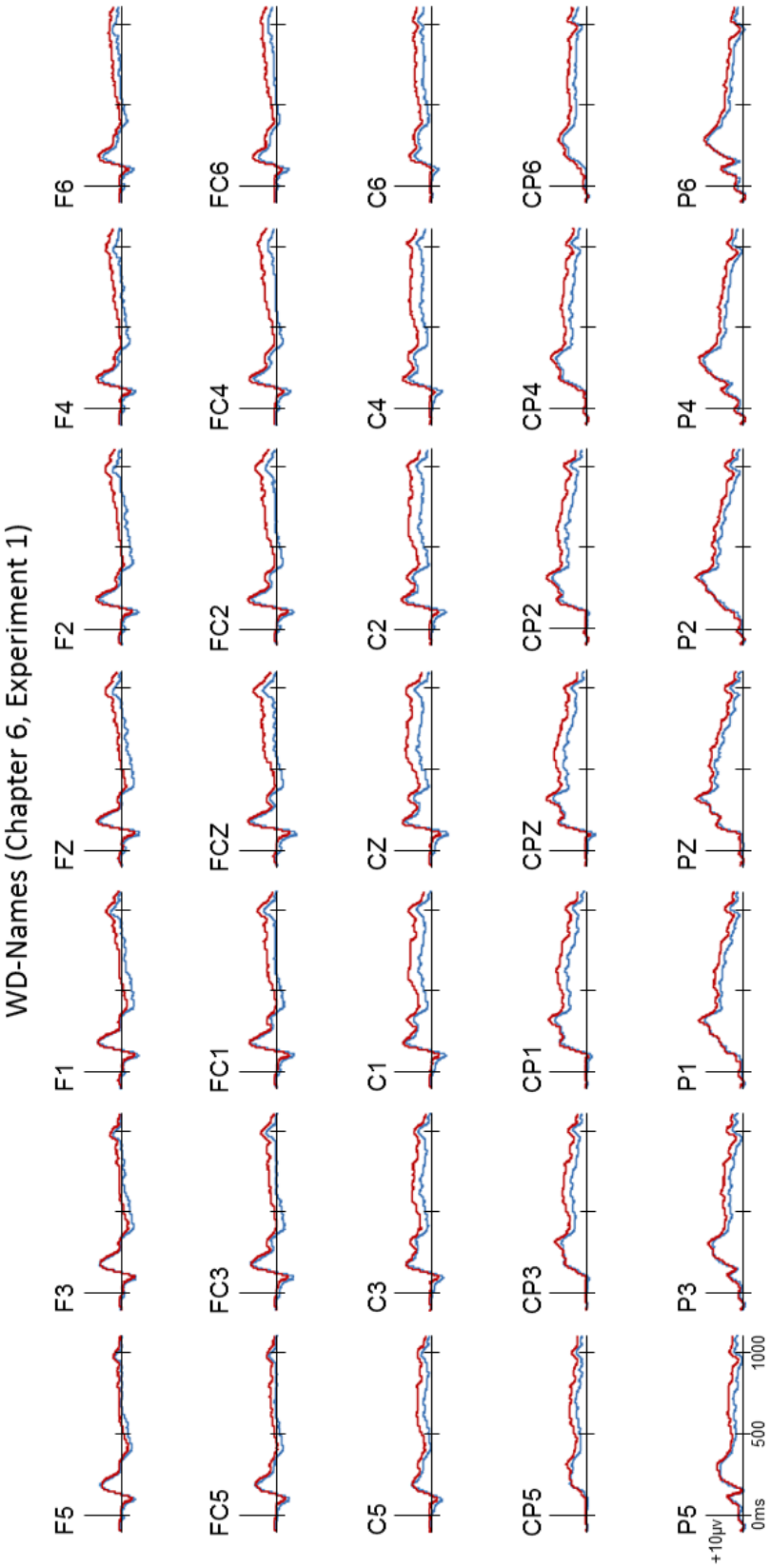


Figure 9.19: Grand average Intact/Rearranged ERPs for WD-Name pairs (Chapter 5, Experiment 1). Rearranged = blue, intact = red.

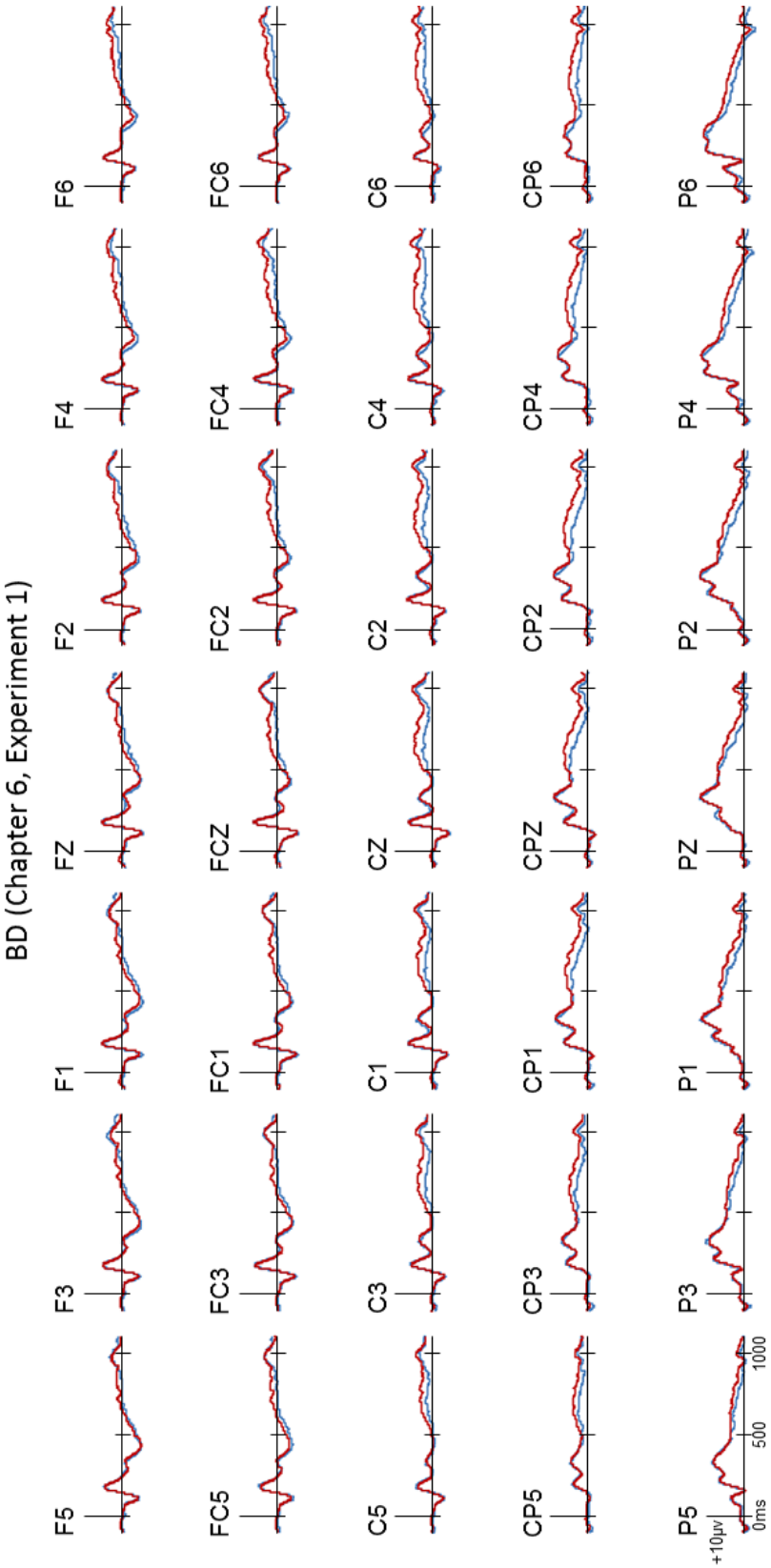


Figure 9.20: Grand average Intact/Rearranged ERPs for BD pairs (Chapter 5, Experiment 1). Rearranged = blue, intact = red.

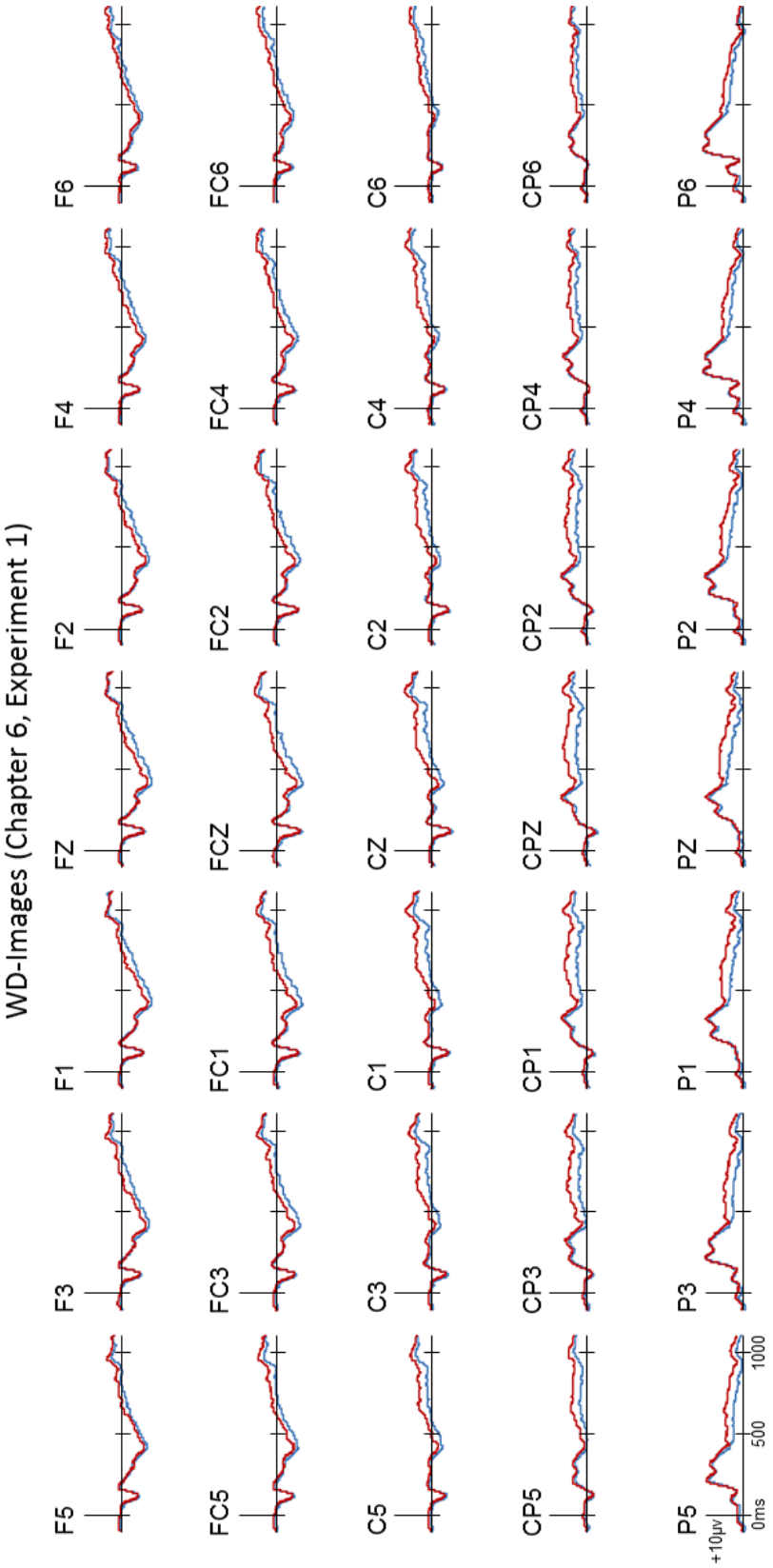


Figure 9.21: Grand average Intact/Rearranged ERPs for WD-Image pairs (Chapter 5, Experiment 1). Rearranged = blue, intact = red.

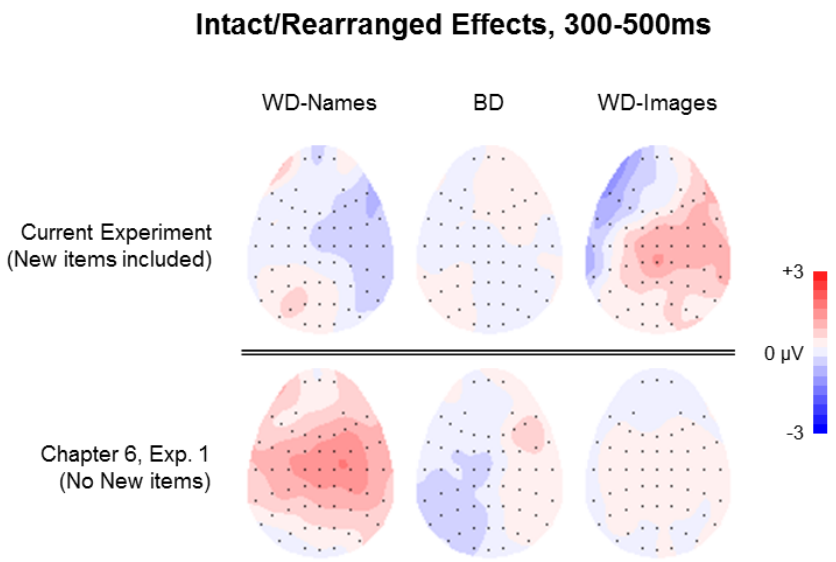


Figure 9.22: Topographic distribution of intact/rearranged effects between 300-500ms. The effects are shown separately for each pair type (*WD-Names*, *BD*, *WD-Images*) and for the two experiments.

on that data revealed a significant main effect of intact/rearranged $[F(1,26) = 4.40, p = .046]$, such that ERPs to intact pairs were more positive than to rearranged pairs, but the only interaction approaching significance was a marginal interaction with hemisphere $[F(1,26) = 4.03, p = .055]$. This reflected the fact that the positivity was broadly distributed, but with a slight right hemisphere bias.

9.4.6 Mid (500-800ms) intact/rearranged effects

A topographic summary of the intact/rearranged effects between 500-800ms is provided in Figure 9.23. In contrast to results from the previous time window, repeated measures ANOVA on the current experiment between 500-800ms revealed a significant main effect of intact/rearranged $[F(1,23) = 13.69, p = .001]$, reflecting greater positivity across the scalp for intact than rearranged pairs. An interaction with site $[F(1.09, 24.98) = 10.39, p = .003]$ indicated that this positivity was strongest at superior electrodes. Finally, a four-way interaction between intact/rearranged, pair type, hemisphere and site $[F(2.21, 50.73) = 3.97, p = .022]$ suggested possible differences in the lateral distribution of the effect across pair types, but it was absent once intact/rearranged effects were rescaled; there was

no evidence of topographic differences in the intact/rearranged effect across pair types.

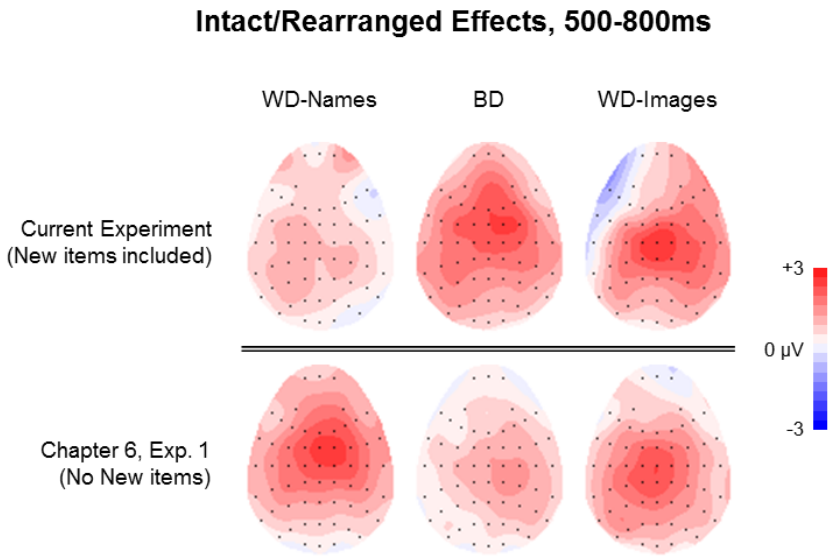


Figure 9.23: Topographic distribution of intact/rearranged effects between 500-800ms. The effects are shown separately for each pair type (*WD-Names*, *BD*, *WD-Images*) and for the two experiments.

Repeated measures ANOVA on the data from Chapter 5 (Experiment 1) similarly revealed a main effect of intact/rearranged $[F(1,26) = 29.05, p < .001]$. This also interacted strongly with site $[F(1.14,29.58) = 15.31, p < .001]$, reflecting a comparable intact/rearranged effect (i.e. positive, with superior focus) to that observed in the current dataset.

9.4.7 Late (800-1100ms) intact/rearranged effects

Late epoch intact/rearranged effects are summarised in Figure 9.24. Repeated measures ANOVA on the current experiment between 800-1100ms also revealed a significant main effect of intact/rearranged $[F(1,23) = 9.64, p = .005]$, reflecting greater positivity across the scalp for intact than rearranged pairs. An interaction with site $[F(1.14,26.25) = 8.35, p = .006]$ indicated that this positivity was strongest at superior electrodes. A three-way interaction between intact/rearranged, location and hemisphere $[F(1.67,38.29) = 4.55, p = .022]$ reflected greater intact/rearranged differences at posterior locations on the left of the scalp, but large differences at both anterior and posterior locations on the

right. Four-way interactions between intact/rearranged, pair type, location and hemisphere $[F(2.70, 62.12) = 3.99, p = .014]$, and intact/rearranged, pair type, hemisphere and site $[F(2.05, 47.12) = 4.70, p = .013]$, suggested possible differences in the distribution of the effect across pair types. Analysis of the rescaled data confirmed the distribution of the intact/rearranged effect differed according to pair type, location and hemisphere $[F(2.65, 60.92) = 3.09, p = .039]$; while all three pair types showed greatest intact/rearranged differences at parietal locations on the left side of the scalp, WD-Image pairs uniquely showed the opposite pattern (greater positivity at frontal locations) on the right side of the scalp. The interaction between pair type, hemisphere and site was no longer significant after rescaling.

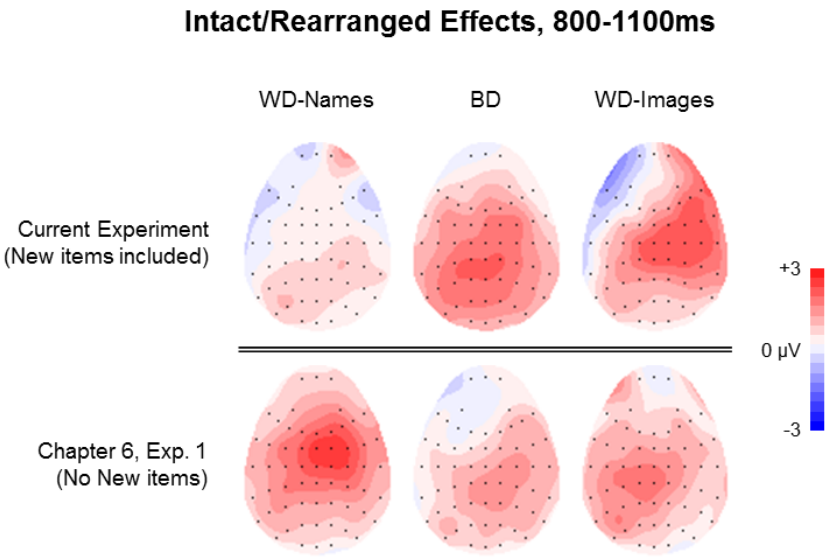


Figure 9.24: Topographic distribution of intact/rearranged effects between 800-1100ms. The effects are shown separately for each pair type (*WD-Names*, *BD*, *WD-Images*) and for the two experiments.

Once again, we repeated the analysis on the data from Chapter 5 (Experiment 1). This revealed a main effect of intact/rearranged $[F(1, 26) = 27.57, p < .001]$ and interaction with site $[F(1.20, 31.09) = 9.57, p = .003]$; a superior positivity to intact compared to rearranged pairs was thus present during 500-800ms and 800-1100ms time windows in both experiments. The three-way interaction observed above, between intact/rearranged, location and hemisphere, was also present in the data from Chapter 5 $[F(1.76, 45.71) = 7.02, p = .003]$, reflecting a parietal focus on the left, but not right hemisphere. A four-way interaction between

intact/rearranged, pair type, location and site [$F(4.37, 113.67) = 3.47, p = .008$] was present after the data was rescaled [$F(4.25, 110.55) = 3.27, p = .013$]. This may have reflected the fact that while all three pair types showed similar parietal intact/rearranged effects, WD-Names had a more superior distribution at frontal sites.

Finally, we checked whether the intact/rearranged difference in later time windows actually reflected response preparation. We divided participants into two groups according to the hand used for each type of response, and included it as a between-subjects factor in a repeated measures ANOVA on the intact/rearranged difference between 800-1100ms with factors of hemisphere[5] and pair type. No main effects or interactions between the three factors were significant for either the current experiment, or Experiment 1 from Chapter 5 (crucial hemisphere by group interactions: $p = .763$ & $p = .195$), meaning that the intact/rearranged effect was not driven by differences in lateralized readiness potentials.

In summary, intact/rearranged effects were broadly similar regardless of whether the task involved an explicit old/new decision (current experiment) or only the discrimination of intact from rearranged pairs (Chapter 5, Experiment 1). In both experiments, intact pairs produced more positive going ERPs than rearranged pairs, starting from around 500ms and lasting beyond the offset of the test stimulus at around 1000ms. The positivity was greatest at superior electrodes throughout, but in the later time window there was also a left-parietal focus to the effect. Intact/rearranged differences during the early time window were limited, although there was some evidence that differences onset earlier when new pairs were not included at test. No significant interactions with pair type were found before the late time window (800-1100ms), at which point some differences at frontal electrodes began to emerge, although these were somewhat inconsistent across the two experiments.

9.5 Discussion

In Section 9.3 the behavioural data revealed qualitative differences across pair types, which interacted with the type of decision being made. Firstly, images were

[5]We averaged the activity from 6 electrodes for each hemisphere, FC1, FC3, C1, C3, CP1 and CP3, in order to capture the relatively centrally located intact/rearranged effect.

less confidently rejected as new than they were accepted as old, while name-name pairs, and to some extent between-domain pairs, were rejected and accepted with comparable confidence (and higher overall confidence than image pairs). This might reflect the comparative distinctiveness of individual names compared to images, within each full dataset. In other words, new images were more easily confused for previously-encountered ones. This makes some sense if we suppose that representations of images are less well separated from each other than representations of names are: the presentation of a new image with some similar properties to those that were previously encountered may prompt full or partial reactivation of old images, increasing false alarms (equivalent to the reduction in mean confidence to new pairs observed).

In contrast, for the intact/rearranged decision it was between-domain pairs which showed differences relative to the two within-domain conditions. For between-domain pairs, intact pairs were identified more confidently than rearranged pairs were, but this pattern was not present in either within-domain condition. Rearranged decisions were made with comparable confidence regardless of domain, but intact pairs were recognised more confidently for between than within-domain conditions (Figure 9.6(b)). In this case, then, we must consider an explanation that accounts for this more confident recognition, rather than a reduction of confidence as suggested for new image pairs above. The pattern is, for example, consistent with unitization in the sense that recognition for the encoded (intact) pair is selectively improved.

Analysis using the DPMSD model provides us with further information about this pattern, ascribing it to more frequent recollection to intact than rearranged between-domain pairs. This conflicts somewhat with an interpretation of unitization which predicts item-like recognition for unitized pairs, since there was no evidence of recognition in the absence of recollection. The data suggest that recollection was triggered unusually frequently by presentation of an intact between-domain pair, which may reflect stronger or more easily reactivated associations between the components, or greater availability of holistic information.

9.5.1 Including new items in an associative recognition test

Here we examined associative recognition under two different conditions, either with or without a concurrent old/new decision. There was some evidence that intact and rearranged pairs dissociated earlier in the neural record when new items were not employed, but this was not reflected in shorter reaction times - in fact the opposite pattern was observed (Figure 9.7). Given that participants were not pressured to respond quickly (and were in fact prevented from responding before the onset of the response screen), early differences in the neural record would not necessarily be expected to correlate strongly with response times. Broadly speaking, the differences in intact/rearranged contrasts across the experiments were small, and provided little evidence that participants employed very different strategies or qualitatively different processing to distinguish intact from rearranged pairs when new items were also included.

9.5.2 The FN400 and familiarity

In the current experiment, we found clear evidence of early frontal positivity to old pairs compared to pairs of new items, consistent with the presence of FN400 effects. This pattern has been observed in many experiments, and generally linked to familiarity (e.g. Nessler et al., 2001, Trott et al., 1999, but see also Curran, 2004) though others have argued instead that it may reflect conceptual priming (e.g. Olichney et al., 2000, Voss and Paller, 2006). Both explanations are arguably consistent with our data, and in particular with the estimates from the DPMSD model for both this experiment and that carried out in Chapter 6. These studies suggested that familiarity (or conceptual priming) provided a means of distinguishing old from new pairs, and the FN400 was accordingly present in old/new contrasts. We note, however, that like others (Curran et al., 2002) we find the FN400 to be present even for images, when conceptual information is presumably minimal, perhaps better supporting its conceptualization as an index of familiarity or broader priming effects, than specifically conceptual priming.

Regardless of its precise interpretation, the most important finding is that FN400 effects were not present in intact/rearranged contrasts, which is consistent with evidence from the DPMSD model that familiarity did not contribute to in-

tact/rearranged decisions. Assuming the FN400 does reflect familiarity or priming, the electrophysiological data provides a vindication of the DPMSD model's characterization of familiarity and recollection. Equally, it supports the conclusion that discrimination of intact from rearranged pairs, even on the basis of global characteristics enhanced by unitization, is not supported by familiarity.

9.5.3 The left-parietal old/new effect and recollection

There was evidence of greater old/new positivity between 500-800ms over parietal electrodes on the left, but not the right, of the scalp. This is consistent with the left-parietal old/new effect, which has generally been linked to recollection (Paller and Kutas, 1992, Rugg and Yonelinas, 2003, Smith, 1993), and which we would expect to appear in old/new contrasts in this study since recollection was - according to the DPMSD model parameters - key to associative recognition.

The size of the left-parietal old/new effect has been argued by some to reflect the amount of information recollected from an episode (Vilberg et al., 2006, Wilding, 2000). This is not a claim we are able to verify with this experiment. Images were, according to the DPMSD model, less frequently recollected as intact or rearranged than either name-name or between-domain pairs, but left-parietal old/new effects did not differ across pair type, with the only differences occurring at the front of the scalp. It is possible, however, that the lower frequency of recollection to intact pairs was compensated by the retrieval of quantitatively more information. Image pairs were associated with relatively low recollection strength, however, so if more information was retrieved on these trials it was not generally sufficient to distinguish between intact and rearranged pairs. It should be stressed that this interpretation is plausible: information recollected for image-image pairs may have been diagnostic of old/new (thus causing the left-parietal old/new effect to appear in the old/new ERP contrasts) while at the same time not being strongly diagnostic of intact/rearranged (thus not appearing in the DPMSD estimates of recollection). Such a pattern was arguably observed in Chapter 6, where the frequency of recollection to individual images was higher than that for names, but the opposite pattern was observed for recollection of associative information. If this is the case it suggests that individual images might be frequently (though not especially accurately) recalled, even while their associations prove harder to

recollect. Future studies should take care to separately measure the strength and rate of recollection (as estimated behaviourally) for different tasks, in combination with imaging data, to better understand how recollection of different types and quantities of information might contribute to a given memory decision (and how these relate to putative imaging correlates of 'recollection'). It may be that recollection is dissociable not only in terms of its strength and rate, but also in terms of the type of information it retrieves (e.g. item v associative).

Interestingly, the left-parietal old/new effect did not appear to vary across pair type, and appeared whether the stimuli were meaningful (names) or not (abstract images). These data are therefore difficult to reconcile with the argument that the left-parietal old/new effect reflect recollection of meaningful information, while anterior effects - which were also observed across all three pair types - reflected recollection of meaningless stimuli (Cycowicz and Friedman, 2007, Galli and Otten, 2011, MacKenzie and Donaldson, 2007, Yick and Wilding, 2008). Perhaps instead the distribution of old/new effects reflects qualitative recollection differences of another kind, correlated in that study with the meaningfulness of stimuli, but the fact that both anterior and left-parietal effects are matched across the three pair types here makes our data less than ideal for teasing apart the differences between the two. Instead, we may be better able to interpret the ERPs associated with recollection using the contrast between recollection of intact and rearranged pairs.

Did the size of the left-parietal old/new effect vary with test condition, i.e. whether the old items were part of an intact or rearranged pair? The size of the effect did nominally appear to be greater for intact than rearranged pairs, but this was, in the main, not localised to left-parietal electrodes in the traditional 500-800ms time window. There was evidence that the left-parietal effect was greater for intact than rearranged pairs in the later, 800-1100ms epoch however, and this was present in both studies. It is important to note that rearranged pairs showed much greater negativity across the scalp, particularly at central and parietal locations, and this effect is somewhat difficult to disentangle from any possible differences in left-parietal old/new effect size. Significant interactions with location and hemisphere in the intact/rearranged contrast do, however, support what appears to be the case from the topographic maps in Figure 9.24, namely that late epoch intact/rearranged differences included a left - but not right - parietal

increase. More broadly, however, the DPMSD model estimates suggest that intact pairs were more frequently recollected than rearranged pairs in Chapter 5 (Experiment 1), but not in the current experiment, making it difficult to ascribe the greater positivity to intact pairs as being necessarily related to recollection frequency.

The left-parietal old/new effect has been previously shown to be larger for internally generated than externally presented stimuli (Senkfor et al., 2002, Wilding and Rugg, 1997), raising the possibility that internal sources are more frequently or strongly recollected than external sources (Hicks et al., 2002). Larger left-parietal old/new effects for intact than rearranged pairs might be explained within this framework if rearranged pairs were mainly recognised by recollecting originally paired components (an external source) but intact pairs sometimes prompted recollection of other, internally generated information. Arguably, larger left-parietal effect sizes for meaningful than meaningless stimuli (Cycowicz and Friedman, 2007) might reflect a greater availability of internally generated information, assuming that meaningful stimuli in that studies were more easily manipulated at study and used to generate source information.

In summary, at most the current data could be said to be consistent with the view that the left-parietal old/new effect indexes the amount of information recollected. Perhaps instead the left-parietal old/new effect (at least in this later time window) primarily reflects recollection of information that is present at test, which would arguably be greater for intact pairs since there is a greater study/test overlap than for rearranged pairs. Taking this argument further, it could be suggested that one such source of evidence is the (internally generated) global pair information that might support recognition of intact pairs, i.e. the left-parietal old/new effect might index retrieval of unitized or holistic information, and not cued recall of separate (externally generated) study items that were associated with rearranged pair components. If this is the case, might other components of the ERPs conversely reflect cued recall? One candidate is the late posterior negativity, or LPN.

9.5.4 The LPN and response preparation

Comparisons between correct old and new responses revealed clear LPN effects in later epochs, which differed in size according to whether pairs were intact or rearranged. As noted earlier, there are a number of possible interpretations for what the LPN reflects. One possibility is that it is related to response preparation, perhaps directly reflecting activity in motor and premotor cortex (Kuo and Van Petten, 2006). While possible, we consider this explanation unlikely for two main reasons. Firstly, the lateral position of the effect did not vary with response hand, as would be expected if it was evoked by preparation of a particular motor response. Thus, if the difference represents preparation of a response, it must reflect processes occurring prior to the preparation or movement of a particular hand. Secondly, physical responses were temporally separated from presentation of test stimuli in both experiments. Stimuli were shown for 1000ms, followed by a further 500 or 1000ms blank screen, and then a response screen at which point participants were free to make a button press (buttons were disabled until this response screen and participants were additionally instructed not to make a response before it). Thus all responses were made at least 1500ms after stimulus onset, and on average around 2500-3000ms post-stimulus (Figure 9.7). The relatively long gap between response screen onset and button press (around 1000-1500ms on average) makes it unlikely that motor responses were frequently selected and prepared some 1000ms *before* response screen onset, i.e. 1500-2000ms before the response was made, yet this corresponds roughly to the late epoch in which the LPN was strongly observed (800-1100ms post-stimulus).

Alternatively, the LPN may be associated with the selection or monitoring of an intact/rearranged/new judgment, but not the preparation of motor functions required to make the physical response itself (Mecklinger, 2000). The difference between test conditions may reflect differences in the time taken to reach a decision: as is clear from Figure 9.7 rearranged responses were made some 250ms later than intact or new responses, and rearranged responses were also associated with larger magnitude LPNs than were intact pairs. Perhaps, therefore, decision making or response monitoring processes were concurrent for intact and new pairs (leading to smaller differences in the components underlying the LPN) while these processes occurred later for rearranged pairs (leading to larger differences in the components). This explanation is more plausible than a manipulation of

the lateralized readiness potential, since the processes need not be directly related to motor activity, which they appear not to be. Once again, however, if responses were selected in advance of the response screen onset, we might expect that they should be made rapidly when the response screen appeared, whereas in fact they were made some 1000-1500ms later. More importantly, while rearranged pairs exhibited both greater LPN and RT differences with new pairs, compared to intact pairs, there was no evidence that the size of the LPN correlated with reaction time differences within either intact [$r(70) = -.10, p = .415$] or rearranged [$r(70) = .06, p = .603$] pairs. Thus, the correlation between reaction time and LPN magnitude does not constitute strong evidence that the two are directly linked. It is plausible that extended reaction times and larger LPN magnitude for rearranged pairs are downstream effects of a third variable, especially given the absence of a correlation once test condition is accounted for.

9.5.5 The LPN and episodic retrieval

It is worth noting in any case that an appeal to response selection does not explain why rearranged pairs were associated with longer response times than intact pairs. The simplest account is that rearranged pairs prompted some additional or longer processing than intact pairs (or that the engagement of longer or additional processing led more frequently to rearranged responses), and that this processing was responsible for increasing both reaction time and LPN amplitude. Such an account is consistent with the broad view that the LPN reflects additional episodic reconstruction or retrieval (Johansson and Mecklinger, 2003).

Some researchers have further suggested that the LPN might reflect systematic, as opposed to heuristic, decision processes (Johnson et al., 1993, Leynes and Phillips, 2008). Heuristic and systematic decision processes are proposed to use different types of information to judge the relative likelihood of a particular source or context (e.g. the image a name is paired with at test, meaning the pair is intact) compared to that of a different source (e.g. a different image, meaning that the pair is rearranged). Heuristic processes rely on rapid evaluation of how well the details retrieved from memory, such as a particular colour, match each source. Systematic processes are more controlled and might rely instead on reasoning and directed search, for example by recalling an internally generated association

between the two items at study.

Our results are compatible with this view in the sense that rearranged pairs might be identified based on systematic search processes more frequently than intact pairs. For example, recalling an original study pairing in order to reject a rearranged test pair involves a systematic search and retrieval of an episode, and its relationship to the current stimuli. Beyond this, however, greater retrieval of heuristic information might result from unitization, especially given that intact pairs showed lower LPNs but greater left-parietal positivity, and therefore possibly greater reliance on heuristic rather than systematic processes. If global or holistic information was available for intact pairs, this should provide additional details which could be recollected and (heuristically) compared to the test pair, increasing the relative reliance on heuristic information for intact compared to rearranged pairs.

The LPN did not appreciably differ across pairs, suggesting that its function is not stimulus-specific, or at least was not modulated by the presence or otherwise of lexical stimuli in this study. In particular, it is hard to argue that it reflects the quantity of specifically perceptual information retrieved (Cycowicz and Friedman, 2003, Cycowicz et al., 2001, Wolk et al., 2007), since it was of equivalent size in conditions which differed strongly in the amount of perceptual detail likely to have been retrieved (names v images).

9.5.6 ERP differences across pair types

We can draw some tentative conclusions about how the retrieval of between-domain pairs might differ from the retrieval of names or images. Despite small differences, all three pair types showed broadly similar ERPs. While this may at first seem surprising, given the large and consistent differences in performance across pair types observed in the behavioural data, these would not necessarily be expected to yield corresponding differences in ERPs which are formed from (selected) correct responses. Instead, therefore, these data suggest that differences in performance across pair types are not driven by differences in the relative contribution of processes indexed by FN400 (familiarity), LPONE (recollection) or LPN (episodic reconstruction). It is important to note that this does not preclude qualitative differences in memory across pair types, only that those retrieval pro-

cesses linked to the old/new effects listed above did not appear to differ by pair type.

The majority of differences across pair types appeared to relate to lateral distribution over frontal electrodes, and were present for rearranged pairs in early epochs, and intact pairs in later epochs. In particular, intact between-domain pairs gave rise to greater frontal positivity, particularly on the left of the scalp, compared to within-domain pairs. If these do reflect genuine differences in neural activity across the pair types, their location might suggest they relate to frontal lobe activity. The three pair types showed no difference in terms of the ERP effects commonly linked to recollection (LPONE and LPN), suggesting that while between-domain pairs were more frequently recollected than expected, this was not based on a qualitative difference in the retrieval processes indexed by these effects. Perhaps the qualitative differences between them were driven not by the way information was retrieved, but instead by how it was interrogated or used.

9.5.7 Summary

The ERPs examined here were broadly similar across the two experiments, suggesting that the inclusion of new pairs did not have a major effect on the cognitive processes engaged in discriminating intact from rearranged pairs. The data also support the main contention drawn from the DPMSD model estimates throughout this thesis, namely that familiarity did not contribute to associative recognition (but it did support old/new recognition). The putative correlate of familiarity, the FN400, was present in contrasts between old and new pairs, but not between intact and rearranged pairs. The data also reveal evidence of a LPONE to both intact and rearranged pairs, consistent with the importance of recollection for the task. The LPONE was not modulated by pair type, and thus did not appear to be material-specific. Finally, an LPN was clearly present. The LPN was strongest for rearranged pairs but did not vary with response hand or pair type, suggesting that it was more likely to be related to episodic reconstruction than to the retrieval of perceptual details specifically, or to response preparation. Importantly, the fact that it was greater for rearranged than intact pairs (while the FN400 did not differ) supports the conclusion from Chapters 5–7 that intact/rearranged discrimination may be supported by probabilistic recollection - but not continuous

familiarity - of holistic pair properties.

Chapter 10

General Discussion

In this thesis we set out to describe the experience of human episodic memory in terms of its underlying processes. In Chapter 4 we carried out two experiments with the aim of characterizing recollection as either a thresholded or continuous phenomenon. We found strong evidence that recollection was thresholded and also graded, and that neither of these properties could be explained simply by factors at encoding, as has previously been suggested (DeCarlo, 2003, Mickes et al., 2010).

This characterization of recollection is of interest in itself, but it also has implications for the nature of both the neural signal underlying recollection and the cognitive processes which support it, including how they might fail or be impaired by aging or disease. We shall turn to those later in this discussion, but in this thesis we focussed on one particular implication: how should recollection be modelled? The two most widely used models of memory strength respectively describe recollection as thresholded but not graded (the DPSD model; Yonelinas, 1994) or as graded and continuous (the UVSD model; Green and Swets, 1966). The results from Chapter 4 strongly suggest that both of these models are fundamentally flawed in the way they treat recollection, and instead support the use of another class of models, in which recollection is explicitly both graded and thresholded. Such models have been argued for previously, generally on the basis of superior fits to confidence data (Onyper et al., 2010), but the experiments described in Chapter 4 provide particularly crucial evidence since they constitute, as far as we are aware, the first direct test of the underlying assumptions of these

models.

10.1 The DPMSD model of episodic memory

In the introduction to this thesis, we outlined three criteria for a good model of episodic memory. Such a model should be parsimonious, it should fit the extant data well and it should be grounded in theory so that its parameters can be linked clearly to properties of episodic memory. A model which passes all three of these criteria should, as a result, be capable both of explaining existing patterns of data in terms of memory processes and also of producing testable, sensible predictions for future studies to address. The DPMSD model passes all three of these criteria, and thereby constitutes a useful framework to guide the investigation of human memory.

Firstly, the model provides a good fit to the data gathered in this thesis. In the four confidence-based experiments carried out in Chapters 5–9, the DPMSD model fit recognition data for both associative and non-associative tasks well. In one case a simpler VRDP model provided a more parsimonious account, and in another the even simpler UVSD model described the pattern of data most efficiently. Thus, sometimes, the DPMSD model is less parsimonious than some of its alternatives. Crucially, however, neither of these more parsimonious models are capable of explaining the full range of data found in this thesis, and thus such findings are likely better accounted for by a lack of statistical power than a drastic over-interpretation of data by the DPMSD model. Perhaps most crucially, the DPMSD model ascribes a clear psychological meaning to each parameter, allowing predictions to be readily made. This further allows for the model to be tested in ways other than the examination of confidence data, which we argue may be ill-suited to this particular purpose (see Section 2.3.5).

10.1.1 Alternatives to the DPMSD model

We have argued that the DPMSD model is a better explanation of the underlying data than existing accounts, but we also note that it is not especially parsimonious. Is this lack of parsimony justified (beyond the empirical justification

implied by likelihood ratio tests on its parameters)? To answer this question we need to consider how well more parsimonious models can account for the same data. The DPSD model cannot easily account for source and associative data in which memory strength is shown to be variable (e.g. Mickes et al., 2010, Slotnick, 2010) and the underlying assumption in the model which is responsible for this - i.e. that recollection does not result in measurably variable confidence - has been widely rejected, including by proponents of the model (Yonelinas et al., 2010). The continuous UVSD model conflicts with strong evidence in Chapter 4 which demonstrates that recollection is thresholded. Furthermore, the UVSD model has in any case been frequently shown to be unable to account for associative or source task data without including a thresholded component (e.g. DeCarlo, 2003, Kelley and Wixted, 2001, Onyper et al., 2010, Slotnick, 2010, Slotnick and Dodson, 2005, Wixted et al., 2010). The simple UVSD model is parsimonious, but it is an inadequate model of memory strength since it cannot account for the greater-than-expected rates of very strong and very weak recollection, which the DPMSD model ascribes to thresholded recollection.

A potentially parsimonious alternative model, which does not require an explicit threshold to account for such a pattern, has also been suggested by Shimamura and Wickens (2009). This hierarchical relational binding theory (hRBT) model is based on the suggestion that while recollection and familiarity are not clearly and qualitatively different types of memory response, there is a certain nonlinearity about memory strengths in the sense that most memories will be weak (and associated with little recollected detail) but those which are not weak will tend to be considerably stronger (and associated with a great deal of recollection). The model captures this effect by convolving the traditional Gaussian distribution of strength with an exponential distribution, i.e. it employs an ex-Gaussian distribution which is skewed to the right (higher strengths). More broadly, one could achieve a similar effect by using any skewed Gaussian distribution in place of the symmetric distributions commonly assumed.

Could the confidence data in this thesis be better explained by a skewed distribution of memory strength, such as the hRBT model proposed in Shimamura and Wickens (2009)? Here we highlight two patterns we would expect to observe if it were. Firstly, the long tail of the ex-Gaussian means that a mixture model approximation to it would see greater variance in the stronger distribution than

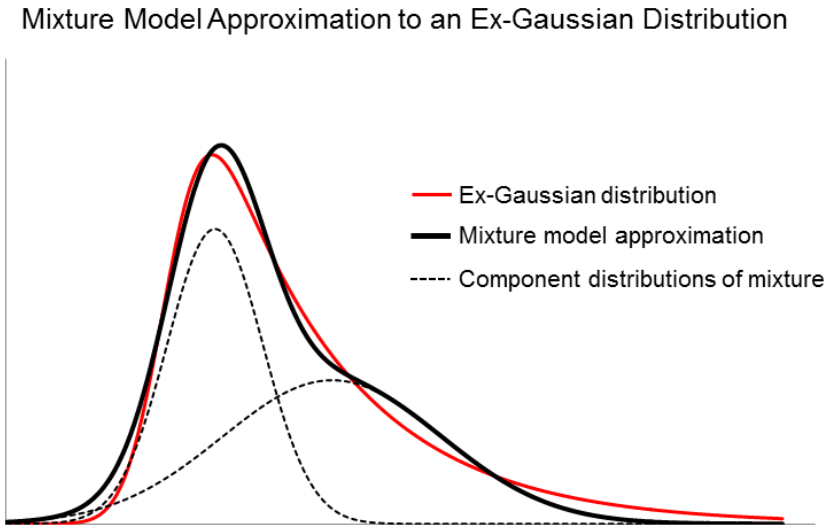


Figure 10.1: Mixture model approximation to the ex-Gaussian. Since the ex-Gaussian distribution is skewed to the right, in order to approximate its shape the stronger distribution in the mixture must have a greater variance than the weaker distribution (here $\sigma^2_{strong}/\sigma^2_{weak} = 5.2$).

the weaker distribution, Figure 10.1. As we discuss later in Section 10.2, however, the mixture model fit to the observed data shows the opposite pattern. We also note later that there are reasons to be cautious when comparing variance across populations with different means, and so it remains possible that in fact a skewed Gaussian distribution provides a good account of memory strength (but not confidence).

Even in this case, other patterns exist in the data which cannot be so easily explained by a skewed Gaussian description of memory strength. As illustrated in Figure 10.2, underlying data that is distributed as an ex-Gaussian should also lead to positively correlated recollection rate and strength when estimated using a mixture of Gaussian distributions. That is, as performance increases and the distribution skews further to the right, both the mean of the stronger distribution and the proportion of trials lying within it should increase. In fact, we (Chapter 6) and others (Onyper et al., 2010) observed the opposite pattern: Names were more strongly, but less frequently, recognised than images. Perhaps these two experiments were unusual, and recollection strength and frequency are not dissociable; perhaps instead the patterns observed might be better accounted for by changes in bias to non-recollected stimuli (see Section 7.4.1). In that

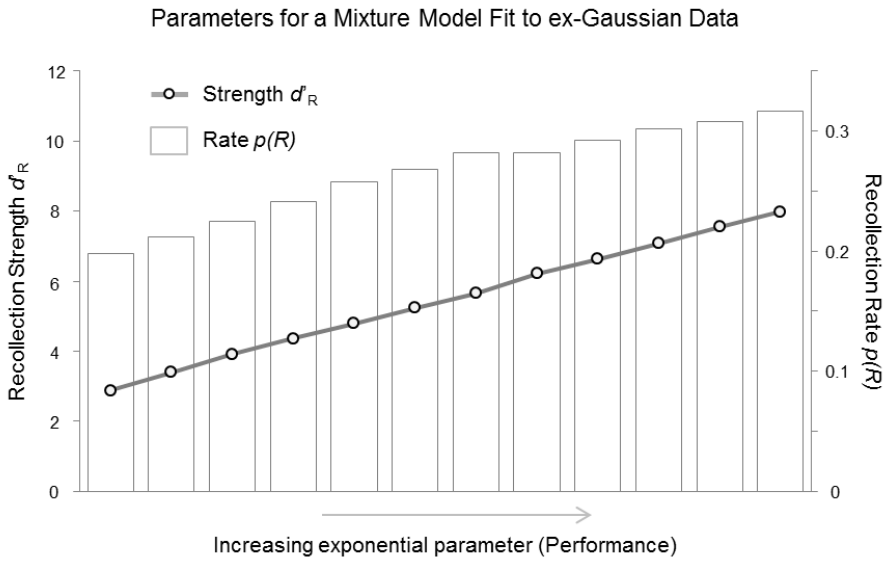


Figure 10.2: Parameters for a Mixture Model Fit to ex-Gaussian Data. If data is distributed as an ex-Gaussian, then as performance (exponential parameter) changes, the rate and strength of recollection in a mixture model approximation should move together. In fact, the two parameters are functionally separable (Chapter 6).

case a model such as hRBT which accounts for the strength and frequency of recollection (or strong memories) with a single parameter may provide a more accurate account than the DPMSD model. In Section 10.4, however, we discuss some possible implications if the strength and frequency of recollection are, to some extent, independent properties.

10.1.2 The nature of recollection: Areas of agreement

An important point to note is that, despite the many arguments in the literature, the most successful measurement models (UVSD, DPSD, DPMSD, hRBT) share some fundamental properties. Most notably, all describe unstudied items using a Gaussian distribution of memory strength, but describe studied items using a Gaussian distribution that is amended in some way: increased variance (UVSD), a high threshold (DPSD), additional Gaussian distributions (DPMSD) or a positive skew (hRBT). Each amendment has the result of increasing the variance and, with the exception of the UVSD model, the skew of the target distribution. As a result, all but the UVSD model imply an uneven distribution of memory strength, such that a subset of trials is remembered with much greater confidence than most.

The DPSD and DPMSD models explicitly link this subset to recollection, and the remainder to non-recollected trials. Thus, for these models, the threshold parameter describes the proportion of memories which are retrieved together with their context, and supposedly provides a quantitative index of this contextual retrieval from simple behavioural data. The hRBT model makes the link with recollection and contextual retrieval less clear cut, but no less real. In this case the long tail of strong memories is accompanied by 'recollective experience' for these trials (Shimamura, 2010), and the extent to which this occurs is controlled by the rate parameter of an exponential distribution. Thus across these three models the nonlinearities present in ROC data are all ascribed to the increasing influence of recollection. The UVSD model is less clearly linked to recollection and familiarity, but it has been argued that the greater target variance may stem from the combination of recollection and familiarity for these items (Wixted, 2007a), and further that higher-strength memories should generally include more recollected details (Wixted et al., 2010). As previously discussed in Section 2.3.1, the lack of a precise functional interpretation of the UVSD model should, we believe, be regarded as a major weakness.

We would argue, however, that the theoretical interpretations of the DPSD, DPMSD and hRBT models are clear, and further that the differences between them in this respect are relatively small: all three agree quite closely on the reason for target distribution nonlinearities, and differ instead in how these nonlinearities should be described mathematically. This is not to say that such differences are unimportant. Indeed, as we have argued (Section 2.3.5), and demonstrated in practice (Section 5.5), accurate modelling of parameters can be important, especially where they are used to identify quantitative relationships between cognition and other data, such as from functional imaging, or neurocomputational models. However, most of the major competing models concur that the degree to which a target distribution diverges from a simple Gaussian reflects the relative engagement of processes supporting recollection, or at least the availability of relational, episodic information. Thus, the parameters in each model which capture this divergence can be used as (imperfect) indices of qualitative differences in memory retrieval. The results in this thesis support the argument that of the three models, the DPMSD model provides the most accurate measure of this engagement in most circumstances.

10.1.3 Limitations of the DPMSD model

As with all models, the DPMSD model is an imperfect simplification, and it is important to consider what boundary conditions might apply to it. Specifically, the DPMSD model assumes only two or three (depending on the task) distinct types of responses. For example, in an item recognition task responses to targets are either recollected or non-recollected (familiar), and these are represented by separate Gaussian distributions in the mixture. We could easily, however, imagine other classes of responses, representing factors such as partial attention, false recollection, motor error or recollection of multiple study episodes, and a fully specified model might incorporate many more Gaussian distributions for each of these and other factors. Thus, in the parsimonious DPMSD model all of these different factors will be incorporated into the two target distributions: recollected or familiar. As a result, it should not be immediately assumed that the parameters represent recollection, familiarity or guessing are process-pure estimates. Prudent consideration of the possible types of response which could account for each distribution in the mixture, combined with careful task design, are important factors in correctly interpreting responses using a DPMSD model.

A related point is that when different distributions are poorly separated, the DPMSD model is less useful. This can occur for different reasons: perhaps recollection and familiarity do not provide sufficiently different levels of strength, recollection is extremely frequent or infrequent, or confidence ratings introduce too much noise. For example, computational models suggest that high stimulus overlap can lead to both weaker recollection, and less distinction from familiarity in terms of distribution shape (Norman and O'Reilly, 2003). In all of these cases the resulting overlap of the two underlying distributions may be such that they approximate a single distribution (e.g. see Figure 2.9). One effect of this is that the UVSD model may be accepted as the most parsimonious description of the data, providing an apparent basis for this model to be (erroneously) declared a correspondingly appropriate model of the underlying mechanism (e.g. Wixted, 2007a). A further effect is that parameter estimates from the DPMSD model will become decreasingly precise, informative and accurate as the extent of overlap increases.

An example of this problem can be found in Chapter 9, which serves to high-

light the behaviour of the DPMSD model expected when the underlying target strengths better approximate a single Gaussian distribution. When fit to the item recognition data (discrimination of pairs of old items from pairs of new items), likelihood ratio tests reject the inclusion of a separate familiarity parameter and produce estimates of recollection rate which are near ceiling. The model thus becomes a close approximation of the UVSD model - it comprises a target distribution in which virtually all trials (around 97%) fall into a single Gaussian distribution, with the few remaining trials approximately following the lure distribution. We discussed the reasons for this in more detail in the preceding chapter; briefly, the pattern arises because of a task design which allows many different bases for memory strength. Thus, while we stress that a superior (i.e. more parsimonious) fit to confidence data for a UVSD than a DPMSD model does not necessarily validate UVSD as a good account of the underlying memory processes, it may yet serve as a useful test of whether the data is sufficiently multimodal to safely interpret the DPMSD model parameters.

Finally, the DPMSD model is, as we noted in the previous section, less parsimonious than some of the alternatives: simpler unidimensional accounts, such as the EVSD or UVSD models, will often provide more parsimonious fits according to fit criteria such as AIC, and especially BIC. As well as the importance of careful task design which we stress above, another key requirement when drawing important conclusions from the DPMSD model is therefore sufficient data. Previous analyses have recommended upwards of 50 points per condition to obtain stable ROC parameter estimates for even parsimonious models (Macmillan et al., 2004), and the DPMSD model might require more since a greater number of parameters must be estimated. While to some extent the stability of the model can be improved by collecting more finely grained confidence ratings (i.e. more than the standard 6 points, thereby retaining more information about the distribution, see Chapter 4) and by averaging parameter estimates across large numbers of participants, this may not completely overcome the problem when the number of responses is much lower than 50, as we saw in Chapter 7. In sum, the DPMSD model is useful, but it requires a sufficient number of datapoints to provide stable parameter estimates, and these parameters should in all cases be interpreted carefully. We note of course that these particular limitations are also true for other models of memory strength, which the DPMSD model has a

number of key advantages over. One of these advantages, which we shall discuss in the next section, is that it explicitly models recollection as both thresholded and graded, which allow the variance in recollection strength to be measured.

10.2 Patterns in confidence rating variance

The relative variance of targets and lures has been an important piece of evidence for models of memory to account for (Glanzer et al., 1999, Green and Swets, 1966, Ratcliff et al., 1994). The results in this thesis cast light on why targets are associated with higher variances than lures, and highlight how the DPMSD and similar models can be used to address this question as well as others - such as how recollection and familiarity interact in episodic recognition.

10.2.1 Patterns of variance in this thesis

If recollection and familiarity are both continuous variables, which are combined to form a single evidence value on each trial, then target distributions should form a single distribution with greater strength variance than lures. If recollection is continuous but absent on some trials simply because information was not attended to or encoded, then (after removing guesses) targets should still have a higher variance than lure distributions. Alternatively, if recollection is essentially a thresholded process, it should become increasingly diagnostic as the memory task it supports becomes easier. For example, recollection should be somewhat diagnostic on a source task, more diagnostic on an associative recognition task, and highly diagnostic on an item recognition task, where even weak recollection should yield high confidence that the item was previously encountered. Thus, confidence strength for recollection-based decisions should be less variable for easier tasks, since the majority of responses will give very high confidence for the task (even though they may differ greatly in terms of underlying memory strength).

In Table 10.1 we summarise the standard deviation ratios found in this thesis, across different tasks. It shows that we consistently found non-recollected trials to vary in confidence more than recollected trials, and that this ratio differed

according to the task being performed. Commonly, targets in recognition studies are found to have greater variance than lures, prompting the development of, and providing continued support for, the UVSD model (Green and Swets, 1966, Wixted, 2007a). At first glance, therefore, our findings seems at odds not only with the UVSD model and its dual-process explanation of how such variance differences arise (by combining evidence), but also with the extant data itself.

Chapter	Task	Mean s.d. ratio
6	Item Recognition	0.717
5	Associative Recognition	0.761*
6	Associative Recognition	0.733
7	Associative Recognition	0.769
9	Associative Recognition	0.767
4	Source Recall	1.334
4	Source Recall	1.436

Table 10.1: Mean recollected v non-recollected standard deviation ratios by task. Trials for which recollection was successful were less variable in terms of reported confidence than those for which recollection was absent, except where the strength of recollection was crucial to the task (source recall). *Inclusion of the variance ratio did not significantly improve the fit of the data in Chapter 5.

Closer consideration reveals that the variance data in this thesis are not only compatible with the existing literature, but potentially illuminating as to the reason it arises. Here we highlight the variance of trials for which the DPMSD model calculates recollection to have been successful. The fact that these trials have lower variance than non-recollected trials implies that *separate* rather than complementary information may be being used to judge confidence when recollection occurs. Indeed, if we continue the assumption that greater variance in confidence may arise from the combination of different sources of evidence (Wixted, 2007a), these data suggest that in fact it is non-recollected trials which are less process-pure, and are based on widely varying (and possibly non-diagnostic) evidence.

Recollection variance also seems to vary systematically with the task being performed, and this is presumably a function of how diagnostic weak recollection is for a particular decision. For example, if even weak recollection provides relatively good evidence that an item has been encountered before, most recollected trials will be associated with high confidence, resulting in smaller variance for these trials. In such cases high-threshold accounts such as the DPSD model will more closely approximate the data. However, if the strength of recollection is critical to a decision, such as in source tasks where it is more important to recollect an item or association accurately, the variance in strength will be reflected in more variable confidence judgments.

Regardless of the relationship between recollected and non-recollected targets, when (all) targets are compared to lures, they are more variable in strength, and this is similar to the pattern observed in the literature. For example, in the single item recognition decision from Chapter 6 the mean target/lure standard deviation ratio is 1.23, which is very close to the value of 1.25 frequently quoted as being representative of item recognition studies (Glanzer et al., 1999, Ratcliff et al., 1994, Wixted, 2007a).

10.2.2 The interaction of recollection and familiarity

The discovery that confidence ratings to targets were more variable than for lures prompted a re-evaluation both of the common EVSD model of memory responses, and theoretical memory models which predicted matched target and lure variances (Glanzer et al., 1999). Theories were developed to explain why this variance pattern should be the case: perhaps different amounts of memory strength were added at study (Hilford et al., 2002, Wixted, 2007a), or else different sources of evidence - perhaps from qualitatively distinct memory processes - were combined at test (Mickes et al., 2009, Yonelinas et al., 1998). While the former explanation has been apparently rejected by an empirical test of its predictions (Koen and Yonelinas, 2010), it is difficult to disprove the latter by simply examining the overall variance of targets, nor to determine whether different sources of evidence are combined on each trial. In particular, an open question in dual-process theory is how recollection and familiarity are related: whether they are effectively independent, as some models assume (Jacoby, 1991), or whether

familiarity is always available when recollection occurs (Joordens and Merikle, 1993). While double dissociations between recollection and familiarity have led some to declare the two are functionally independent (see Section 1.2), this does not exclude the possibility that familiarity and recollection are highly correlated. Indeed, it seems likely that this should be the case since the strength of each depends upon, for example, study duration (Murdock, 1974). Furthermore, the fact that each process is functionally independent also allows that their evidence might be combined in order to make a decision (Wixted, 2007a). Under the assumption that recollection is thresholded but variable, this would lead to a prediction that non-recollected targets (familiarity only) should have lower variance in memory evidence than recollected targets (familiarity plus recollection).

Here, by virtue of rejecting the UVSD model and instead using a more detailed DPMSD model, it is possible to gain a clearer view of why targets exhibit greater variance: it is because a proportion of them are recollected, and associated with high memory strength, while the remainder are not. Importantly, the model reveals that while multiple sources of evidence (such as familiarity and recollection) may both contribute over the course of a task, this does not necessarily mean that they are combined on a trial-to-trial basis. We can go further, and state that according to the results in this thesis, some trials are process-pure in the sense that recollection is not successful or diagnostic, and therefore that these trials rely on familiarity (or perhaps other sources of non-mnemonic evidence). It is, in fact, intriguing that recollected trials are associated with relatively low variance in confidence, since this further suggests that decisions may in those cases also be comparatively process-pure, and reliant upon recollection, either alone or predominantly. This may of course simply be the product of our particular choice of stimuli, task and instructions, yet if it were to be established in general, as has the overall increase in variance for targets, it would support the view that, when it is successful, recollection should provide the sole or primary basis of recognition decisions (Yonelinas, 1994).

We do note that this does not necessarily imply that other evidence is unavailable, only that it is not used to make a confidence rating. Perhaps the strength of familiarity can in future be measured in the presence and absence of successful recollection, for example by examining the accuracy of a speeded forced-choice recognition decision conditional on whether recollection was later detected. Fi-

nally, we also note that there are important, and often underappreciated constraints on interpreting variance differences between sets of confidence ratings. The most important of these is that the scale being used by participants, or the relationship between memory strength and confidence, may not be linear, and therefore that variances for populations with different means cannot be compared (Egan, 1975, Rouder et al., 2010). Perhaps by counterbalancing where on a scale (e.g. centre vs edge, or lower end vs upper end) each type of decision should be placed, or by examining variances in more grounded or less subjective metrics such as accuracy, stronger conclusions can be drawn from examining the comparative spread of memory responses. Such problems are limited mainly by difficulties in task design, though it is should also be noted that reasonable assumptions about the mapping of psychological variables to physical responses are required to draw strong conclusions about latent memory strength, as opposed to confidence responses (Rouder et al., 2010). Importantly, a model such as DPMSD which allows the properties of recollected and non-recollected responses to be separately investigated - but not a UVSD model which collapses the two in the interests of parsimony - provides both an explanation for existing patterns in memory data, and also a tool for investigating these and other patterns in more detail.

10.3 When does a recollection threshold occur?

We have argued above that on the basis of evidence from this thesis, the DPMSD model is a useful and accurate account of memory strengths, and that it achieves this by incorporating a thresholded, graded process with a continuous, graded one. Here we consider why a threshold might arise in memory strength data, and in particular we pose the question: When does it arise? During encoding, storage or retrieval, or some combination of these?

10.3.1 Thresholds introduced at encoding

As we noted in the introduction, the same pattern that according to the DPMSD model reflects thresholded recollection could be accounted for by other interpretations. These ascribe the existence of a threshold to factors other than recollection

failure (DeCarlo, 2002; 2003, Hautus et al., 2008, Mickes et al., 2010, Slotnick, 2010). While we rule out the theory put forward in most of those articles, namely that the threshold is determined *entirely* by processes at encoding, it remains possible that on some trials recollection fails to occur because the information required was not encoded.

For example, some proportion of trials at encoding might not be attended to (DeCarlo, 2003). When presented at test, these trials would be essentially new to the participant and memory strength would correspondingly be described by the lure distribution, even though they are characterised by the experimenter as old. More broadly, and perhaps more subtly, a similar effect has been argued to occur even when items are attended to (Mickes et al., 2010, Wixted, 2007a). By this argument, something akin to a true memory threshold exists at encoding. Some minimum level/length of attention or encoding effort may be required in order to encode certain kinds of information, particularly relational and context information that later rely primarily on recollection. When a required level of encoding is not reached then the relevant information cannot be later recollected, leading once again to a threshold in the data. Regardless of the particular explanation invoked, the important theoretical point of both accounts is that this threshold is related to the presence or absence of information at the point of retrieval, as a consequence of whether or not it was encoded.

10.3.2 Thresholds introduced after encoding

According to the results from Chapter 4, however, a threshold does exist that is not simply a consequence of encoding. Perhaps this may result during retention (or the transfer of memory from shorter-term to longer-term storage): information may be lost subsequent to encoding, for example due to catastrophic interference, and thereby become impossible to later retrieve. Alternatively, a retrieval process which is critical to recollection might be inherently thresholded, rendering recollection success probabilistic regardless of the strength of stored representations. If this is the case, recollection may fail even when the information being sought is present in memory. It is important to note that the three possibilities are not mutually exclusive, for example a thresholded retrieval process does not preclude the possibility that some trials could have simply failed to be encoded

or else lost entirely during retention. In addition, it seems likely that the probability of retrieval should rely to a large extent on the integrity of the information being retrieved, which in turn depends upon the strength of encoding and level of interference during retention. Whether there is indeed such a link between encoding and the *frequency* of recollection, as opposed to its strength, is yet to be empirically demonstrated as far as we are aware. More generally, an important aim for future research should be to determine how experimental manipulations shown to affect recollection do so: by affecting its strength or its frequency.

One practical consequence of a retrieval, rather than encoding, threshold for recollection is to be found in the imaging approaches used to isolate its neural substrate. Often, fMRI is recorded during encoding of stimuli and then sorted according to the results of a later test phase, undertaken outside the MRI scanner. If the ability to later recollect is primarily determined at study, for example by a thresholded pattern of encoding as suggested by Kelley and Wixted (2001), the contrast between later-remembered and later-forgotten items at study should be strong, and reflect the areas of the brain which play an active role in successful encoding. Alternatively, if as some results in this thesis suggest, the threshold emerges mainly as a result of probabilistic processes underpinning retrieval, the contrast may be less informative - for example some later-forgotten trials may have been strongly encoded. In this case, fMRI contrasts measured at retrieval might better highlight areas in the brain associated with the success of processes that are critical to episodic recollection. These brain areas would be of particular interest to researchers investigating how memory declines with age and disease.

10.3.3 A threshold at retrieval?

It is important to stress that the questions of when and why a threshold emerges in recollection are important directions for future research, and strong conclusions cannot be made about, for example, the presence of a threshold at retrieval per se on the basis of the data in this thesis. To be clear, the critical result that the recollection rate decreased as a function of study-test delay (Chapter 4, Experiment 2) does rule out an entirely encoding-determined threshold, but does not necessarily require that a threshold exists at retrieval. It is possible that a subset of memory traces (i.e. those later not recollected) could be lost

or made inaccessible over the retention interval, or they might drop below some minimum integrity as a result of interference[1]. Nonetheless, as we outlined in Chapter 4, there does exist some evidence that neural activity prior to retrieval is predictive of successful recollection (Herron and Rugg, 2003), which arguably suggests that a retrieval threshold may exist which is functionally independent of how the memory trace is encoded and stored.

Furthermore, our data provide an important constraint in the question of whether a threshold is produced at retrieval or during storage of a memory, and it is that the threshold appears to act specifically on one type of memory retrieval (recollection) but not others (e.g. familiarity). According to the behavioural data from Chapter 6 and supported by electrophysiological data (and arguably also behavioural data, but see Section 9.3.1), old items can always be somewhat differentiated from new items by familiarity, or some other continuous source of evidence, such as priming. This is in contrast to recollection, which provides no information on some trials. Thus even when recollection fails to retrieve a particular episodic detail or association, at least some information about the episode remains which can be used to recognise the previously-encountered stimuli.

How might a retention threshold account for this? Perhaps associative or source information is stored separately to item information, in networks which are more vulnerable to the kind of catastrophic interference suggested above. Computational models of neural networks arguably support such a possibility, since the capacity of such networks (i.e. the number of patterns that can be stored and later fully retrieved by pattern completion) is subject to a limit, determined by the number of neurons and connections in the network (Hertz et al., 1991). The implication here is that different types of information are somehow distinguished at encoding, and stored differently depending on their properties. One proposal is that such a distinction occurs naturally as a function of stimulus complexity (Cowell et al., 2010). This account supposes that as the visual system processes information of increasing complexity (edges and simple shapes in early visual areas such as V1 and V2, more complex objects further downstream), there develops a similar hierarchy of stored memory traces along the ventral visual stream and medial temporal lobe cortex. The model has so far only been developed and

[1]Note that in this case the trace is so corrupted as to be completely inaccessible to recollection-supporting processes, not simply noisy.

tested on item recognition of varying complexity, although the authors stress that they expect it should extend naturally to the hippocampus, where complex associations and interactions between objects might be stored. Critically, however, if this or similar models are to account for a storage threshold, they would do so by invoking functional differences in the networks used to store different types of information: specifically, those networks storing associative information should have a limited capacity and be vulnerable to catastrophic interference, while those storing less complex item information should be more robust.

An alternative view, consistent with a retrieval threshold, is an accumulator model in which a graded neural signal triggers a thresholded retrieval response (Donaldson et al., 2010, Ratcliff, 1978). By this account, evidence of oldness may accumulate in parietal cortical regions, and an 'old' decision is made when the weighted evidence acquired reaches a sufficient level (resulting in a 'moment of recognition'). Only memories which accumulate enough evidence to exceed the required threshold will be recollected, and they will vary in accuracy and perceived strength because of their differing amounts of evidence. In support of this type of account, different brain regions have been shown to either track with the amount of mnemonic evidence available, or 'switch on' in a binary fashion as a decision is reached (Ploran et al., 2007).

10.3.4 Strategies for determining when a threshold arises

It may be possible to confirm whether a threshold is a consequence of retention or retrieval using similar logic to that used in Chapter 4. If a retrieval process is genuinely thresholded it should be possible for a participant to demonstrate successful recollection on some trials for which recollection has previously failed, without re-encoding between tests, or to manipulate the probability of recollection (not just its strength) by changing conditions at test but not prior to it.

If a threshold does occur at retrieval, might we be able to isolate the process which fails? In Chapter 9 we found a late posterior negativity (LPN) in ERP old/new contrasts, which may reflect the engagement of a subset of processes supporting episodic reconstruction Johansson and Mecklinger (2003). It is possible that the LPN reflects a directed search for episodic information, but not necessarily its success (Friedman et al., 2005). This account is consistent with our findings

that rearranged pairs were linked to larger LPNs and response times than intact pairs, if we suppose that sometimes intact pairs were recognised based on holistic information, while rearranged pairs were more often subjected to a search for episodic details, and relied more on strategic processing (Rotello and Heit, 2000). Perhaps the search processes engaged in recall, which may be reflected by the LPN, could be prone to failure - in which case declining retrieval in older adults might reflect a reduction in successful engagement of these processes, hence poorer recall-to-reject (Healy et al., 2005). It would be interesting to determine whether low-confidence responses (for which recollection presumably failed more frequently) lead to LPNs of a smaller size than high-confidence (recollected) responses did. If so, the LPN may (possibly indirectly) reflect the engagement of a thresholded process, i.e. it would be a correlate of successful recollection. Alternatively, if the LPN does not vary with recollection success, it might instead index non-thresholded processes supporting search attempts, isolating the threshold to factors further downstream in the processing which supports recollection. Using the same logic, we can investigate how other ERP and fMRI correlates vary with recollection success to further fractionate the phenomenon of episodic retrieval, and isolate the critical cognitive processes and brain areas that fail during retrieval. Such findings could potentially provide important areas of research focus with significant long-term benefits to the understanding of why memory declines - and fails - with aging and disease.

10.3.5 Non-mnemonic 'Recollection'

The fact that recollection is thresholded may also have wider cognitive implications. Throughout this thesis we have focused entirely on episodic memory. As we briefly acknowledged in Section 1.1, however, a sharp separation of episodic from other forms of memory is unlikely to be valid. Recollection or a similar set of processes may retrieve not only episodic but also semantic information: does the failure to bring a particular word to mind reflect very different cognitive processing to the failure to recollect an episode from memory? Furthermore, the accuracy of episodic memories is far from perfect, especially in the long term. When a memory becomes contaminated by others - perhaps even by events that we did not experience first-hand (Roediger and McDermott, 1995) - is it still episodic? Can recollection in these circumstance be described as episodic retrieval, or do

recollection and its underlying processes reflect more general cognitive mechanisms for retrieving information? In this latter case, perhaps the thresholded nature ascribed to episodic recollection in fact arises from a more general underlying process, which is simply most noticeable or obvious in the case of memory retrieval failure.

10.4 Dissociating recollection frequency and strength

Regardless of the reason for a threshold, the fact that it is apparently specific to recollection licenses the use of dual-process models which can estimate the contribution of recollection against a background of continuous evidence, such as familiarity or priming. Uniquely amongst these models, however, DPMSD further allows the strength of recollection and the frequency with which it occurs to be separately examined. As noted in Section 10.1.1, these properties seem to be dissociable: in Chapter 6 we reported that while names were more strongly recollected than images, images were actually recollected more frequently (Figure 6.9(b)).

We repeat the important caveat that parameter estimates should always be interpreted with caution: for example this dissociation might be better reflected by a single parameter in an alternative model (though as we have highlighted above, it does not appear to be one such as the skewed distribution suggested by Shimamura and Wickens, 2009). We have justified our use of the DPMSD model on factors other than the utility of its parameters elsewhere in this thesis, primarily in Chapter 4. The most compelling evidence for such a dissociation in general might in future be to test a direct prediction; namely that certain properties of a stimulus, such as perceptual complexity, novelty or semantic richness, should have different effects on recollection strength and frequency. Some data relevant to these questions can be found in (Onyper et al., 2010), which compared recognition of travel scenes to that of words using a VRDP model, which is similar to the DPMSD model we used here. In Experiment 2 of that paper, the authors found a dissociation between recollection strength and rate (in Experiment 1 they observed the same pattern, but it was non-significant). In that paper, travel scenes were recollected less frequently, but more strongly, than words.

10.4.1 How do stimulus properties affect recollection?

A starting point for future studies should be to examine the overlap in properties between travel scenes and names, which were associated with stronger recollection, versus the properties shared by our abstract images and the words used by Onyper et al., which led to weaker but more frequent recollection. Immediately some obvious candidates are ruled out: modality appears not to be decisive, since words and names show opposite patterns, and prior experience or absolute familiarity with a stimulus does not easily explain the pattern either, since it would require that travel scenes are associated with greater pre-experimental familiarity than words. Perhaps the contextual richness of the stimulus might determine the strength of recollection, though this relies on a similarly contentious assumption that in (Onyper et al., 2010) travel scenes were richer in context than words. Alternatively, other factors, such as the distinctiveness of stimuli within a class, might explain the results found here and by Onyper et al., or else have additional divergent or complimentary effects on the rate or strength of recollection.

10.4.2 Neurocomputational and psychological evidence for an effect of stimulus overlap

Distinctiveness is a particularly interesting candidate, given that stimulus overlap was found to reduce the strength of successful recollection, and therefore the bimodality of the recollection strength distribution, in a neurocomputational model of recognition (Elfman et al., 2008, Norman, 2010, Norman and O'Reilly, 2003). This has the dual effect of reducing the strength of recollection (because the recollected-target and lure distributions overlap more) and also increasing the proportion of targets which are recollected: the same pattern observed in Chapter 6 and Onyper et al. (2010).

This raises the intriguing possibility that stimulus distinctiveness, in terms of neural representation, may have a positive effect on the strength or diagnosticity of successful recollection, but conversely make that recollection less likely to occur. In other words, stimuli differ in the extent to which their activations overlap and interact with each other when they are represented in the hippocampus, such that names in Chapter 6 may be more 'isolated', i.e. have less overlap with each

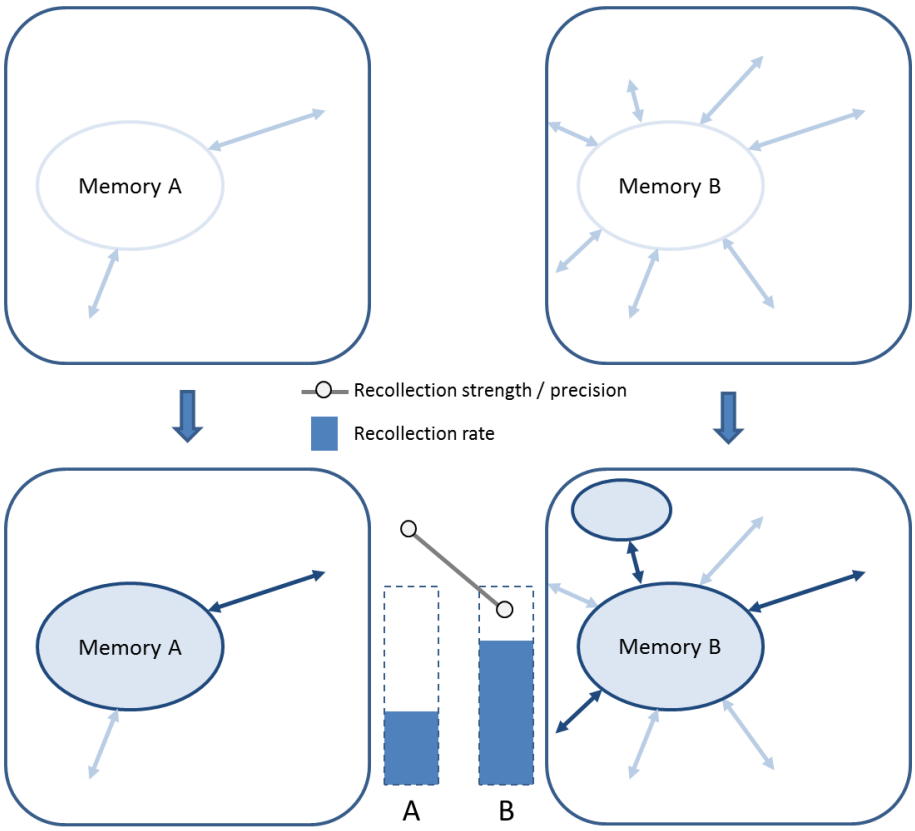


Figure 10.3: Recollection rate and strength as a function of representational overlap. Here two hypothetical representations of previously-encountered stimuli (A and B) differ in the amount they overlap with other representations. Memory A is stored as a relatively distinct pattern of activity, and has little overlap with other representations (light blue arrows). This makes re-activation (recollection) infrequent since few 'routes' of activation are available. Memory B, in contrast, has more overlap with other representations, making it more likely that a search process among these will reactivate Memory B (increasing the rate of recollection). Conversely, however, when reactivation occurs it is more likely to be accompanied by activation of other representations in the case of Memory B, and as a result recollection provides less precise information and lower confidence that Memory B is correct (reducing the 'strength' of recollection).

other or the context of the experiment, than abstract images are. Name representations would thus be comparatively more difficult to re-activate, given a relative sparsity of strong cues at test, but less likely to be co-activated and confused with other representations when re-activation is successful (Figure 10.3). This hypothesis could be tested empirically by using the DPMSD model advocated here, in conjunction with neurocomputational models. A particularly productive approach would entail measuring the accuracy or precision of recollected trials using a similar method to that introduced in Chapter 4, with differing stimulus types (e.g. colours vs locations).

10.4.3 The complementary nature of behavioural and imaging data

We have highlighted here one major advantage that the DPMSD model has over simpler accounts such as the UVSD or DPSD models, besides accuracy: it dissociates episodic recognition further than either of these models, allowing the strength and rate of recollection to be separately measured. In Chapter 9 we used electrophysiological data to examine the processes underlying episodic recognition in even more depth, allowing us to begin to explain associative recognition in more finely grained terms than simply the success or failure of an all-encompassing memory experience such as recollection. Neuroimaging in general is a powerful way of investigating cognition because of the variety and subtlety of changes in brain activity that can be measured, and the temporal resolution of EEG makes it especially powerful for investigating and dissociating the underlying processes for a cognitive task. In this sense, neuroimaging has an important advantage over measuring behavioural responses such as confidence, accuracy or reaction time, and it has greater power to detect qualitative differences across conditions and thereby improve our understanding of memory. Does this mean that behavioural measures such as confidence or accuracy, to which we devote considerable attention in this thesis, are obsolete in all but the most coarsely-grained descriptions of memory?

The answer here is no. Besides their clear practical advantages - the quantity of neuroimaging research that can be done is limited by the number and expense of the equipment they require, as well as the physical demands of the procedure on

participants - simple confidence judgments can provide us with insights that may not be apparent from imaging studies. An important example from this thesis is the characterization of recollection as being thresholded, and the dissociation of recollection frequency and strength. ERPs are formed by averaging trials together, a technique shared by other imaging modalities such as fMRI or MEG, in order to reach the signal-to-noise ratio required to investigate the very small memory-related changes in evoked potentials. This makes trial-to-trial differences - such as the success or failure of recollection - very difficult to infer. Differences in an ERP component may reflect either stronger, or more frequent, engagement of the processes which evoke it. Neuroimaging and behavioural data are best used to complement each other and constrain their respective interpretations, which is the approach taken in this thesis.

10.5 Associative recognition in the context of dual-process theory

We have focused our attention thus far on the nature of recollection. One practical reason why recollection is of particular interest to memory researchers is that it supports episodic associative memory, a fundamental building block of human experience. Given that recollection can fail, and may do so more frequently with age (Howard et al., 2006, Jennings and Jacoby, 1997), it is therefore of considerable interest that recent studies have suggested that familiarity may sometimes also support the recognition of novel associations (Haskins et al., 2008, Mayes et al., 2007, Quamme et al., 2007). In Chapters 5–9 we investigated how recollection and familiarity might contribute to the recognition of novel associations, with particular focus on two current theories: domain dichotomy and unitization.

10.5.1 Domain dichotomy: A viable account?

Domain dichotomy can be considered a stronger theory than unitization, in the sense that familiarity can support recognition of novel pairs in more circumstances, i.e. even when the pair is not unitized. More broadly, pairs of non-unitized items may give rise to associative familiarity, as well as familiarity for

the individual items themselves. We rejected the domain dichotomy theory on the basis that within-domain associations - for which familiarity should contribute to performance - were consistently less easily recognised than between-domain associations, independent of the components from which these pairs were formed. Nonetheless, Mayes et al. 2007 argue that there is some evidence from patient studies which is consistent with domain dichotomy theory (Düzel et al., 2001, Holdstock et al., 2002, Mayes et al., 2004, Vargha-Khadem et al., 1997). There are some weaknesses with both our data and that cited by Mayes et al. 2007: in our case we use a limited number of stimulus classes, and in the case of (Mayes et al., 2007) a limited number of patients (a total of 4 across the studies cited) show the domain dichotomy pattern. Consistent with this latter observation, there are also a number of studies in which patients do not show the domain dichotomy pattern (Stark et al., 2002, Stark and Squire, 2003, Turriziani et al., 2004). Thus it is possible that either the names and abstract images we used were in some way incompatible with the theory, or that the damage to the particular patients studied affected their memory responses in an unanticipated way. One future strategy might therefore be to test within- and between-domain memory in healthy participants (using similar paradigms to those we use in this thesis), but to include a wide range of stimulus classes with the aim of determining whether there is really a consistent effect of within- or between-domain pairings on familiarity or performance overall.

We note, however, that a study cited in support of domain dichotomy healthy participants show the same pattern of performance as we do: between-domain pairs are better recognised than within-domain pairs (Bastin et al., 2010). The same pattern may have manifest itself in another crucial study, which apparently revealed much greater deficits for between-domain than within-domain associative recognition in a hippocampally damaged patient (Mayes et al., 2004). Here 4 within-domain and 18 between-domain associative tasks were reported, and using the authors' own ratings of task difficulty ("Percentage score indicating where between chance and a perfect score the control subjects' mean score fell", Mayes et al., 2004, Table 2, page 767) it can be seen that the within-domain tasks were performed less well on average (47% of a perfect score) than the between-domain tasks were (68% of a perfect score). This does not mean that the tasks were harder *because* stimuli were within-domain: the two types are not directly com-

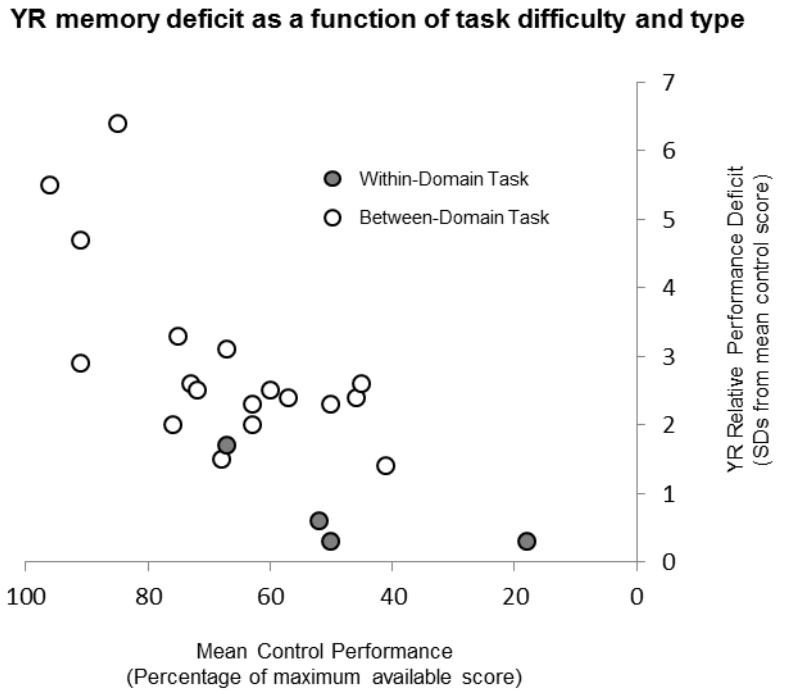


Figure 10.4: Patient YR's associative recognition performance as a function of task difficulty and type. While hippocampally-damaged patient YR shows smaller deficits on within- than between-domain tasks, this may instead be a function of their relative difficulty, since the difference between her performance and that of controls is amplified by high control scores (low task difficulty). Data taken from (Mayes et al., 2004), Tables 3 & 4.

parable since the paradigms used were not consistent. However, it is likely to have an effect on the authors' chosen measure of YR's impairment: her performance relative to the control mean in standard deviations. This is because chance provides a (soft) lower bound for performance, so that if control participants are close to chance, by definition YR's performance could not be considerably worse. This effect can be clearly seen in Figure 10.4, which also shows that control performance was lower for within- than between-domain tasks. Viewed in this way, the results of Mayes et al. 2004 are somewhat less striking than they appear in Figure 3 of Mayes et al. 2007, where task difficulty differences are not visible and only YR's mean z-scores for within- and between-domain tasks are presented.

Thus, perhaps domain dichotomy can be reinterpreted. Rather than stating that within-domain pairs are more recognisable on the basis of familiarity (implying they should be better recognised overall), they may instead be less recognis-

able on the basis of recollection (thus resulting in worse overall performance, as is generally observed). Stated another way, between-domain pairs are benefited by improved recollection relative to within-domain pairs. This is consistent both with the results in this thesis, which suggest the between-domain advantage is based on recollection and not familiarity, and the data from Mayes et al. 2004, which shows better performance by controls on between-domain associative recognition tasks. The patient YR may be unable to benefit from this recollection improvement due to her hippocampal damage, resulting in comparatively greater differences between her performance and that of controls on the tasks in which they were able to use this advantage.

Why might between-domain pairs be better recollected than within-domain pairs? A possible answer to this question can be found in Chapter 6. Here we demonstrated that certain components were better recognised when they were studied as part of a between-domain pair. While in our study this effect was isolated to images, if as some other studies suggest the advantage can be generalised (Criss and Shiffrin, 2004), it might support the observed phenomenon of better recollection of between-domain pairs.

Our results further suggested, however, that component recognition alone did not completely explain the advantage for between-domain pairs; at least for the choice of stimuli in this thesis. Instead, we originally hypothesised that these pairs might be more readily unitized than within-domain pairs. As we note above and in Chapter 5, much of the limited evidence for domain dichotomy might be reinterpreted without recourse to the domain dichotomy theory outlined in Mayes et al. 2007 and reiterated in Montaldi and Mayes 2010. Unitization, however, arguably has a larger supporting literature than domain dichotomy, and we therefore next summarise how our results might be explained in terms of unitization.

10.5.2 Unitization: A viable account?

Some aspects of the results in this thesis might be considered to be consistent with, or even supportive of, unitization of between-domain pairs. It is important to note, however, that two features of the data are difficult to account for using existing models of unitization. Firstly, the components of between-domain pairs

were well recognised; specifically images were more easily recognised than when they were encoded as part of a within-domain pair. In this case, therefore, purported unitization improved recognition for associations without impairing recognition of the components (and possibly improving it). This does not fit easily with a description of unitization which supposes that the individual representations of each component are to some extent replaced by a combined representation, i.e. one which supposes that there should be component-level costs associated with unitization (e.g. Mayes et al., 2007).

The second important feature of the data is that familiarity (or any other graded evidence available in the absence of recollection) did not contribute significantly to associative recognition - a pattern observed consistently in the DPMSD model fits throughout Chapters 5–9. In support of this, electrophysiological effects did not clearly discriminate between intact and rearranged pairs before 500ms in Chapter 9; familiarity is normally accompanied by distinctive neural correlates during this epoch (Curran, 2000), as indeed it was for comparisons between old and new pairs. The availability of familiarity to recognise unitized pairs is generally considered to be one of the defining features of unitization (Diana et al., 2007), yet it is notably absent in these data.

10.5.3 Narrowing the definition of unitization

How, then, might these results be reconciled with the extant literature on unitization? We suggest firstly that the term is too broad and all-encompassing if it is to be used to describe the results in this thesis. Perhaps unitization should be defined specifically in the terms suggested by its name: that component items are perceived and recognised as a single unit, with corresponding costs to the perception and recognition of its individual components. Importantly, retrieval of the unitized components should be equivalent to retrieval of a single item - i.e. unitized representations should be recognised rapidly on the basis of familiarity. The between-domain pairs in this thesis would appear to fall clearly outside of this definition; their recognition was based heavily if not exclusively on recollection and the individual components were well recognised when presented individually.

Nonetheless, the consistent associative recognition advantage for between over within-domain pairs shared many characteristics with the purported effects of

unitization: it was not explained by differences in component recognition, was based largely on more accurate, rapid and confident recognition of intact pairs, and was disproportionately disrupted by altering the overall appearance of the pair at test using a perceptual-switch. In aggregate, these results point towards the use of stored holistic characteristics to recognise intact between-domain pairs, albeit falling short of the stricter definition of unitization outlined above.

10.5.4 Systematic or heuristic recognition of associations?

More speculatively perhaps, intact pairs in general showed reduced LPN amplitude compared to rearranged pairs in Chapter 9; perhaps this reflected reduced dependence on explicit, systematic retrieval of episodic details. This reduced dependence might occur because global or holistic characteristics are sometimes sufficient to support recognition of intact pairs, but absent for rearranged pairs, which instead more frequently prompt additional search processes. To be clear, this would suggest that intact pairs can be recognised based on some global pair characteristics, akin to unitization, but that unlike unitization this information can only be accessed and acted upon when recollection occurs.

It should be noted that while a larger LPN to rearranged than intact pairs is consistent with this account, it does not necessarily require such a qualitative difference between processes involved in recognising the two, being explicable instead simply in terms of greater episodic reconstruction (i.e. from two episodes) for rearranged pairs. If, however, the LPN does reflect systematic search processes in particular, it should be possible to manipulate the size of the LPN for intact pairs. For example, if the perceptual-switch experiment from Chapter 7 were to be repeated while also recording scalp EEG, perceptually-switched intact pairs should show larger LPN amplitudes than non-switched intact pairs, because global, heuristic information would be impoverished and recognition would require more directed, systematic retrieval. In support of this prediction, reaction times to intact pairs in Chapter 7 increased when they were perceptually-switched, but rearranged pairs were unaffected, suggesting that perceptually-switched intact pairs required longer (e.g. directed and systematic) processing.

10.5.5 Evidence for unitization in the wider literature

If such effects can be produced independently of unitization, then unitization (as defined above) can only be established by demonstrating that component recognition is correspondingly impaired, and that recollection is not necessary to support recognition of unitized pairs. This latter requirement must be established carefully, by acknowledging that recollection is graded and models such as DPSD which do not account for this can erroneously ascribe recollection-based performance to familiarity. Equally importantly, remember-know judgments are of little use to detect the *presence* of diagnostic familiarity, as opposed to its relative contribution across conditions where it is diagnostic, since participants may report weak recollection or non-diagnostic familiarity as 'know' responses (Wixted et al., 2010).

Given these requirements, how much strong evidence is there for episodic unitization in studies purported to demonstrate it? In short, the evidence now appears relatively indecisive for non-lexical stimuli. As well as studies which find no evidence of unitization for novel images (e.g. fractals, Speer and Curran, 2007), two of those which found familiarity for word-colour associates (Diana et al., 2008) and combinations of facial features (Yonelinas et al., 1999) rely on the DPSD model for evidence of familiarity. This does not mean that familiarity definitely did not support source and associative memory in these cases, but the ROC data must be reanalysed using a more accurate model before any strong conclusions can be drawn. This is because the same pattern can be explained by more frequent (graded) recollection (Mickes et al., 2010). In support of unitization one study found greater FN400 ERP effects to individual faces whose associate was subsequently recalled (Jäger et al., 2006), compared to those whose associate was not recalled (and this effect was unique to 'unitized' faces, which were pairs of faces perceived by the participant as representing the same person). The paradigm in this case is somewhat different from the associative recognition tasks normally used to identify when unitization has occurred; it is not certain, for example, that the results of Jäger et al. (2006) necessarily reflect 'unitization' as opposed to stronger familiarity for faces that were seen twice at study. Furthermore, ERP effects for faces are difficult to interpret in general, since they may be different to those found in studies using words or images (MacKenzie and Donaldson, 2007, Yick and Wilding, 2008, Yovel and Paller, 2004).

The evidence does appear to be stronger but still mixed in the case of word pairs. In general conditions thought to be unitized seem more likely to enhance recognition performance (e.g. Quamme et al., 2007, Rhodes and Donaldson, 2007; 2008; though see Bader et al., 2010, Opitz and Cornell, 2006); speed responses to intact pairs (e.g. Ford et al., 2010, Rhodes and Donaldson, 2007; 2008; though see Bader et al., 2010) and produce FN400, or at least early onsetting, ERP effects (e.g. Bader et al., 2010, Opitz and Cornell, 2006, Rhodes and Donaldson, 2007; 2008; though see Speer and Curran, 2007). It may be that words are uniquely flexible, in the sense that two words can be used to refer to a single unitized concept - either by pre-existing definition in the case of compound words (e.g. Ford et al., 2010) or via generation of a concept at study (e.g. Quamme et al., 2007). The unitization of words certainly seems to lead to qualitative differences in the way that they are later retrieved. Nonetheless, no study has, we believe, satisfactorily demonstrated that unitization definitely invokes familiarity (rather than just better recollection); nor that recognition of individual components is reduced; nor that unitization is common enough in non-lexical stimuli to expect that familiarity should generally contribute to the retrieval of novel associations (Diana et al., 2007, Yonelinas et al., 2010). In fact, the ROC evidence which, according to some, require either continuous recollection or unitization (Diana et al., 2007, Mickes et al., 2010) are entirely accounted for simply by modelling recollection correctly - as graded, and thresholded.

10.5.6 'Unitization' through recollection

In Section 5.3.3 we noted that recognition of novel pairs appeared to be supported by (p.1386):

> ...a process that both looks (Experiment 1), and feels (Experiment 2), like familiarity.
>
> Harlow et al. (2010)

Subsequently, evidence from both electrophysiological and behavioural data suggested that, in fact, recollection was crucial and familiarity did not contribute to associative recognition in any significant sense. Yet a simple dichotomy between recollection and familiarity does not fully explain the pattern of associative recognition that emerges over the following chapters. Perhaps there does exist a qualitative difference between those trials participants reported recollection for,

and those which 'felt like familiarity', but that this difference is related to the type of information recollected from memory - heuristic, holistic characteristics of the pair, versus explicit, systematic search and retrieval of crucial episodic details.

Under this interpretation, also discussed in Chapter 9, evidence for unitization may often be explained in terms of recollection, without recourse to familiarity except in narrow circumstances when words are heavily associated. Heuristic information is available quickly, speeding recognition of intact pairs in general, and especially when attention is paid to elements of the association, such as the global appearance of the pair or a common definition of two words. This information still normally requires recollection, but is sometimes available when explicit associations are not recovered, increasing the overall rate of recollection. This can be misinterpreted as familiarity when remember-know judgments are used, or when the variable nature of recollection is not acknowledged. In short, it seems difficult to conclude from the currently available evidence that either unitization or domain dichotomy allow associations to be recognised on the basis of familiarity in normal circumstances.

10.6 Conclusions, implications and future directions

In (Wixted, 2007a), by way of a thought experiment the author introduces us to a juror in a trial, who serves as an analogy to our own recognition memory (p.169):

> It seems reasonable to suppose that recognition decisions are not process pure in light of compelling evidence suggesting that the recollection process, like the familiarity process, is a graded phenomenon. If both processes are continuous, such that both can be associated with varying degrees of confidence, then not combining them into a single memory signal would be like a juror who does not combine multiple sources of evidence when assessing the defendant's degree of guilt. Although an assessment of guilt could be based on one piece of evidence or the other (e.g., either fingerprint evidence or fiber evidence), combining multiple indicators into the overall assessment would be more efficient, and the same holds true for an assessment of memory strength.
>
> Wixted (2007a)

The argument makes perfect sense, assuming that recollection is, indeed a continuous process. But is it? The nature of recollection is much disputed, but is both central and consequential to the study of recognition memory. In Chapter 4, we have used a novel approach to determine the answer to this critical question. By examining accuracy (instead of confidence ratings) in a source task we have been able to clearly show that recollection is a graded, but thresholded phenomenon. Furthermore, this threshold cannot be attributed simply to attention or encoding failure, as has been previously argued (DeCarlo, 2003). We therefore believe the data presented in Chapter 4 provide compelling evidence that settles a long-standing debate in the field of episodic recognition memory (Wixted, 2007a, Yonelinas and Parks, 2007).

This finding has a number of important consequences. Firstly, it supports the dual-process view of episodic memory, in which recollection can be viewed as functionally distinct from familiarity, and clarifies one of these key distinctions. Beyond this, however, the thresholded and graded characterization of recollection at a behavioural level provides a potential link to computational accounts of the networks and algorithms that may support it. The strengths of recollected memories are distributed in a strikingly similar way to signals based on pattern completion in hippocampally-inspired networks (Norman and O'Reilly, 2003). Thus our data can be considered together with these neurocomputational accounts as convergent evidence for recollection as a form of (unreliable) cued pattern completion, likely operating on networks which store distinct, pattern-separated representations and underpinning our experience of episodic memory

as alternately vivid and frustrating. This is an important avenue for future work to explore, and we have made some speculative suggestions along these lines, potentially relating the amount of representational overlap to qualitative effects on the information later retrieved by recollection. Regardless of whether this particular suggestion has long-term merit, an account of memory which relates changes at a neural level directly to their effects on cognition has huge potential benefits in the understanding of cognitive decline. Quantitative modelling, both of mechanistic neural networks and of their ultimate product at a psychological level, provides the most promising means of achieving this.

Next, the characterization illuminates how recollection should be accurately modelled and we demonstrate that this has a major effect in practice. The two most widely-used models - the DPSD and UVSD models - can each be considered a simplification of the alternative DPMSD model which we advocate here. In both cases the simplification places profound limitations on the usefulness of each model. The DPSD model, which does not treat recollection as graded, dramatically overestimates the contribution of familiarity to recognition. The UVSD model, which does not account for the threshold in recollection, is as a consequence unable to fractionate the contribution of recollection and familiarity at all. Progress in quantitative modelling at a psychological level requires widespread use of the DPMSD model, or refinements thereof.

The DPMSD model has other advantages; in particular it allows recollection to be further fractionated. Recent decades have seen initial discoveries that the medial temporal lobes might be important for episodic memory (Milner et al., 1968) develop into more detailed, complex accounts relating recollection and familiarity to particular brain regions (Ranganath, 2010, Yonelinas, 2002a). In Chapter 9 we used ERPs to fractionate recollection into different subtypes, related to different ERP effects. This approach is well-placed to detect differences in cognitive processing, just as fMRI and PET can be used to identify the neurobiology of memory. Using the DPMSD model, behavioural data can provide a complementary fractionation by separating the strength of recollection from its rate, something which is difficult to achieve using averaged imaging data. Together, these approaches should be combined to advance the field further still, and develop a deeper understanding of how the neurobiology, cognitive processes and functional characteristics of recollection are related.

Finally, the DPMSD model re-emphasizes, through Chapters 5–9, the critical importance of recollection in episodic associative memory, and therefore human experience. We find that accounts which claim familiarity can significantly aid memory for associations may overestimate how strongly or reliably such support can apply, often as a consequence of the way recollection and familiarity have been operationalized and measured. Recollection can be influenced in qualitative and subtle ways, which remain to be fully investigated, but is fundamentally a crucial component of memory.

To continue the analogy we began this section with, suppose one piece of evidence comes from an eyewitness who recognises the defendant, and must determine whether they 'know' them from the scene of the crime or elsewhere. This is an example of associative recognition outside the laboratory: the witness must determine whether the two components, the defendant and the location, co-occurred or were encountered separately. In this situation, simply feeling as though the face and location go together does not, in the absence of recollection, provide strong evidence that they did indeed coincide - and the juror would be well advised not to incorporate this into their decision. Understanding the factors which determine recollection success, and especially its deterioration as we age, should be re-emphasized as a primary aim for future research.

Appendix A

Factor analysis

Statistics derived from memory task data are almost certainly lower in dimensionality than the processes or factors giving rise to this data. For example, a discrimination statistic d' or d_a can quite accurately summarise how successful participants are at separating old from new items, but this success is likely to be determined by a multidimensional set of underlying cognitive factors. One way of separating these latent factors is to model the distribution of memory strengths across responses, as described in Chapter 2.

An alternative approach is to examine how performance correlates across multiple tasks. Factor analysis can then be used to examine the underlying structure of the data, in an attempt to explain the pattern of performance across n different tasks using a smaller number $n-p$ of theoretically meaningful factors. We use this approach in Chapter 6 to examine how recognition of different components and relationships might be supported by latent properties of memory, such as the engagement of different processes or the retrieval of different types of representation.

Generally it is important to be wary about making firm conclusions about the number of factors in a model. Firstly, the proportion of variance explained by a factor will also be a function of how many variables are related to it. For example, if we examine 10 tasks, but only one is hypothesised to use a particular cognitive process of interest, the largest factor representing that process is likely still to explain less than 10% of the data. Furthermore, in cognitive tasks especially the true number of factors underlying the data is plausibly far more than the number

of measured variables, and there is not necessarily any reason to expect that the factors of most interest are amongst the largest, except by careful experimental design which succeeds in emphasising differences related to them. Therefore the selection of factors should be made with reference to existing knowledge or reasonable assumptions about the data. For example, if the variables can be theoretically grouped according to some classification which is thought to have a strong influence on the dependent variable, this grouping should be reflected by a large factor in the model.

We should be similarly cautious about the interpretation of factors as meaningful. Firstly, since the largest factors are those dimensions of the data with the greatest variance, such components should only be considered informative if the signal-to-noise ratio can be assumed to be reasonably high, since otherwise dimensions of high variance may represent primarily noise. Secondly, factor axes can be rotated for a given solution of $n-p$ factors. Different orientations can give rise to different interpretations of each factor in terms of the original variables, and if these are not consistent they provide only weak or suggestive evidence for a particular theory (when not corroborated by other techniques or reasonable assumptions).

In Chapter 6 we use factor analysis to gain an (approximate) insight into the number and nature of latent memory variables which contribute to seven recognition conditions. Given the caveats outlined here, we avoid drawing strong conclusions about the number of significant factors and also explicitly test different rotations of these factors to assess the effect this has on how each factor is interpreted. Most importantly, we make a priori predictions to ground the results and treat the factors as descriptive (e.g. to guide interpretation of model parameter estimates) rather than explanatory in themselves. As part of the analysis we varied the number of factors included in the model as well as their rotation, to test which conclusions were robust and which relied on assumptions about these two parameters.

The following tables summarise the results of these additional factor analyses carried out in Chapter 6. The number of included factors and the type of rotation used are both varied. This allows the effect of assumptions about factor numbers or rotation type to be observed; conclusions drawn in section 6.3.4 are robust against both sets of assumptions.

(a) Sums of squared loadings

	Extraction SSL			VARIMAX SSL			QUARTIMAX SSL		
Component	Total	% Var.	Cum. %	Total	% Var.	Cum. %	Total	% Var.	Cum. %
1	3.213	45.904	45.904	3.184	45.493	45.493	3.198	45.680	45.680
2	1.696	24.222	70.125	1.724	24.633	70.125	1.711	24.445	70.125

% Var = Percentage of total variance explained by factor.

Cum. % = Cumulative percentage of total variance explained.

(b) Rotated component matrices

	VARIMAX		QUARTIMAX	
Condition	1	2	1	2
Name-name pairs	.833	-.035	.831	-.065
Between-domain pairs	.708	.272	.718	.246
Image-image pairs	.605	.401	.620	.379
Names (WD at study)	.923	-.129	.918	-.163
Names (BD at study)	.877	-.091	.874	-.123
Images (WD at study)	-.018	.876	.014	.876
Images (BD at study)	.018	.834	.049	.833

Table A.1: Sums of squared loadings and component weightings for the first two principal components of discrimination across tasks, from section 6.3.4. A total of 70.1% of the variance in discrimination scores across all 7 recognition conditions is accounted for by the first two factors extracted by principal component analysis. The amount of variance explained per component, (a), and the component weightings for each condition, (b), are given for both VARIMAX and QUARTIMAX rotations.

(a) Sums of squared loadings

	Extraction SSL			VARIMAX SSL			QUARTIMAX SSL		
Component	Total	% Var.	Cum. %	Total	% Var.	Cum. %	Total	% Var.	Cum. %
1	3.213	45.904	45.904	2.601	37.156	37.156	3.017	43.101	43.101
2	1.696	24.222	70.125	1.593	22.758	59.914	1.622	23.175	66.276
3	0.704	10.063	80.188	1.419	20.274	80.188	0.974	13.912	80.188

% Var = Percentage of total variance explained by factor.

Cum. % = Cumulative percentage of total variance explained.

(b) Rotated component matrices

	VARIMAX			QUARTIMAX		
Condition	1	2	3	1	2	3
Name-name pairs	.934	.085	.025	.912	.082	-.206
Between-domain pairs	.509	.169	.564	.630	.191	.416
Image-image pairs	.207	.148	.910	.421	.186	.825
Names (WD at study)	.884	-.115	.297	.930	-.106	.078
Names (BD at study)	.794	-.110	.368	.861	-.098	.168
Images (WD at study)	-.103	.844	.220	-.051	.853	.203
Images (BD at study)	.053	.893	.008	.049	.892	-.043

Table A.2: Sums of squared loadings and component weightings for the first three principal components of discrimination across tasks, from section 6.3.4. A total of 70.1% of the variance in discrimination scores across all 7 recognition conditions is accounted for by the first three factors extracted by principal component analysis. The amount of variance explained per component, (a), and the component weightings for each condition, (b), are given for both VARIMAX and QUARTIMAX rotations.

(a) Sums of squared loadings

	Extraction SSL			VARIMAX SSL			QUARTIMAX SSL		
Component	Total	% Var.	Cum. %	Total	% Var.	Cum. %	Total	% Var.	Cum. %
1	3.213	45.904	45.904	2.500	35.717	35.717	2.856	40.797	40.797
2	1.696	24.222	70.125	1.580	22.568	58.285	1.609	22.985	63.782
3	0.704	10.063	80.188	1.109	15.839	74.124	0.926	13.226	77.008
4	0.535	7.648	87.836	0.960	13.712	87.836	0.758	10.828	87.836

% Var = Percentage of total variance explained by factor.

Cum. % = Cumulative percentage of total variance explained.

(b) Rotated component matrices

	VARIMAX				QUARTIMAX			
Condition	1	2	3	4	1	2	3	4
Name-name pairs	.896	.080	-.084	.265	.904	.087	-.201	.147
Between-domain pairs	.370	.105	.232	.864	.504	.138	.180	.801
Image-image pairs	.243	.178	.910	.201	.380	.207	.866	.146
Names (WD at study)	.898	-.091	.272	.138	.937	-.079	.156	.020
Names (BD at study)	.817	-.083	.357	.120	.865	-.070	.250	.011
Images (WD at study)	-.148	.827	.128	.247	-.104	.837	.126	.236
Images (BD at study)	.071	.912	.049	-.090	.057	.910	.017	-.128

Table A.3: Sums of squared loadings and component weightings for the first four principal components of discrimination across tasks, from section 6.3.4. A total of 87.8% of the variance in discrimination scores across all 7 recognition conditions is accounted for by the first two factors extracted by principal component analysis. The amount of variance explained per component, (a), and the component weightings for each condition, (b), are given for both VARIMAX and QUARTIMAX rotations.

Appendix B

Model Recovery

We generated 6-point ROCs for 20 hypothetical participants in an item recognition task (50 targets, 50 lures). Memory strengths were determined using a mixture model, such that both lure and target items were associated with normally distributed memory strength, but targets were more familiar (d'_F=1.2). Additionally, some proportion of targets (λ=0.5) were assumed to be recollected with correspondingly higher memory strength (d'_R=3.6). To form ratings ROCs from the distributions of memory strengths, the 100 points for each participant were binned into 6 ratings as equally as possible (i.e. 16 or 17 trials assigned to each rating). Fits to the UVSD model (5 criteria & 2 memory parameters: strength d' and target variance $v(old)$) were superior to those for the mixture model (5 criteria & 3 memory parameters: familiarity d'_F, recollection rate λ and recollection strength d'_R) according to AIC ($AIC_{Mixture} = 5928$; $AIC_{UVSD} = 5898$). Using a BIC statistic, the UVSD model would have been even more strongly preferred since BIC carries a higher penalty than AIC for additional model parameters.

Performance in this dataset is relatively high ($d_a > 2$). We repeated the analysis to see how reduced performance (equivalently, more inconsistent use of the confidence rating scale) further affects the fit statistics for each model. To do so we added increasing Gaussian noise to the data and recalculated the AIC statistic. As shown in Figure B.1 the UVSD model is increasingly favoured as discrimination declines (either as a result of declining memory performance or greater confidence rating noise, see Figure 2.9). Under realistic task conditions, a thresholded mixture model produces data which is more parsimoniously approximated

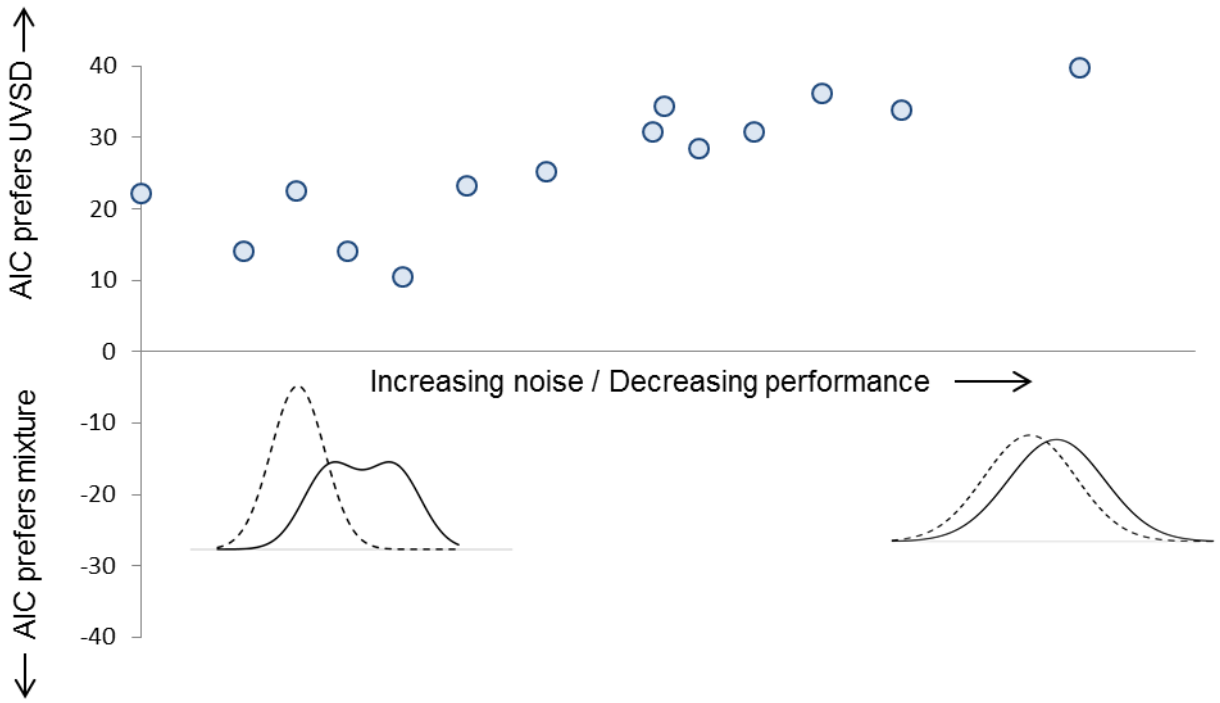


Figure B.1: AIC preference for UVSD model as a function of task difficulty. As performance decreases (equivalently, as the confidence scale is used less consistently) the AIC statistic increasingly prefers the more parsimonious UVSD model of the data to the mixture model which was used to generate it.

by a continuous UVSD model. Despite this, and importantly, parameter estimates for the mixture model were appropriate and close to the true underlying values. For example, the true recollection rate was within 0.7 standard deviations of the estimated value for each of the 14 different sets of simulations plotted in Figure B.1 (average 0.3 standard deviations from the true value), and estimates of recollection and familiarity strength were similarly accurate.

These model recovery analyses are brief, and the appropriateness of different fit statistics will vary with the structure of the dataset being fit as well as the models being compared. It would be too strong an argument to say that AIC/BIC will always select simpler memory models regardless of the underlying pattern of data; at the very least such results suggest a lack of evidence for the more complex, rejected, model. The important point that this simulation illustrates, however, is that absence of evidence does not constitute strong evidence of absence: fit statistics should be used as a guide to *characterise* the data, but one should be wary of using the same statistics to *interpret* the pattern observed.

Bibliography

Aggleton, J. P., McMackin, D., Carpenter, K., Hornak, J., Kapur, N., Halpin, S., Wiles, C. M., Kamel, H., Brennan, P., Carton, S., and Gaffan, D. (2000). Differential cognitive effects of colloid cysts in the third ventricle that spare or compromise the fornix. *Brain*, 123:800–815.

Aggleton, J. P., Vann, S. D., Denby, C., Dix, S., Mayes, A. R., Roberts, N., and Yonelinas, A. P. (2005). Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. *Neuropsychologia*, 43:1810–1823.

Allison, T., Wood, C. C., and McCarthy, G. (1986). The central nervous system. In Coles, M. G. H., Donchin, E., and Porges, S. W., editors, *Psychophysiology: Systems, Processes and Applications.*, pages 5–26. London: Guildford Press.

Anderson, N. D., Craik, F. I. M., and Naveh-Benjamin, M. (1998). The attentional demands of encoding and retrieval in younger and older adults. 1. Evidence from divided attention costs. *Psychology and Aging*, 13:405–423.

Atkinson, R. C. and Juola, J. F. (1974). Search and decision processes in recognition memory. In Krantz, D. H., Atkinson, R. C., Luce, R. D., and Suppes, P., editors, *Contemporary developments in mathematical psychology: Vol. 1. Learning, memory and thinking*, pages 243–293. San Francisco: Freeman.

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed systems and its control processes. In Bower, G. H. and Spence, J. T., editors, *The psychology of learning and motivation: Advances in research and theory.*, volume 2, pages 89–195. New York: Academic Press.

Baddeley, A., Vargha-Khadem, F., and Mishkin, M. (2001). Preserved recognition

in a case of developmental amnesia: Implications for the acquisition of semantic memory? *Journal of Cognitive Neuroscience*, 13:357–369.

Bader, R., Mecklinger, A., Hoppstädter, M., and Meyer, P. (2010). Recognition memory for one-trial-unitized word pairs: Evidence from event-related potentials. *NeuroImage*, 50:772–781.

Bastin, C., Linden, M., Charnallet, A., Denby, C., Montaldi, D., Roberts, J. N., and Andrew, M. (2004). Dissociation between recall and recognition memory performance in an amnesic patient with hippocampal damage following carbon monoxide poisoning. *Neurocase*, 10:330–344.

Bastin, C., Van der Linden, M., Schnakers, C., Montaldi, D., and Mayes, A. R. (2010). The contribution of familiarity to within- and between-domain associative recognition memory: Use of a modified remember/know procedure. *European Journal of Cognitive Psychology*, 22:922–943.

Benjamin, A. S. and Craik, F. I. M. (2001). Parallel effects of aging and time pressure on memory for source: Evidence from the spacing effect. *Memory and Cognition*, 29:691–697.

Bernbach, H. A. (1967). Decision processes in memory. *Psychological Review*, 74:462–480.

Billock, V. A. and Tsou, B. H. (2011). To honor Fechner and obey Stevens: Relationships between psychophysical and neural nonlinearities. *Psychological Bulletin*, 137:1–18.

Bishop, K. I. and Curran, H. V. (1995). Psychopharmacological analysis of implicit and explicit memory: A study with lorazepam and the benzodiazepine antagonist flumazenil. *Psychopharmacology*, 121:267–278.

Bogacz, R. and Brown, M. W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, 13:494–524.

Bowles, B., Crupi, C., Mirsattari, S. M., Pigott, S. E., Parrent, A. G., Pruessner, J. C., Yonelinas, A. P., and Köhler, S. (2007). Impaired familiarity with preserved recollection after anterior temporal-lobe resection that spares the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, 104:16382–16387.

Bowles, B., Crupi, C., Pigott, S., Parrent, A., Wiebe, S., Janzen, L., and Köhler, S. (2010). Double dissociation of selective recollection and familiarity impairments following two different surgical treatments for temporal-lobe epilepsy. *Neuropsychologia*, 48:2640–2647.

Bröder, A. and Schütz, J. (2009). Recognition ROCs are curvilinear - or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35:587–606.

Brown, M. W. and Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, 2:51–61.

Brown, S. and Heathcote, A. (2003). Averaging learning curves across and within participants. *Behaviour Research Methods, Instruments & Computers*, 35:11–21.

Buckner, R. L. and Wheeler, M. E. (2001). The cognitive neuroscience of remembering. *Nature Reviews Neuroscience*, 2:624–634.

Cameron, T. E. and Hockley, W. E. (2000). The revelation effect for item and associative recognition: Familiarity versus recollection. *Memory and Cognition*, 28:176–183.

Chatrian, G. E., Lettich, E., and Nelson, P. L. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity. *American Journal of EEG Technology*, 25:83–92.

Cipolotti, L., Bird, C., Good, T., Macmanus, D., Rudge, P., and Shallice, T. (2006). Recollection and familiarity in dense hippocampal amnesia: A case study. *Neuropsychologia*, 44:489–506.

Cipolotti, L., Shallice, T., Chan, D., Fox, N., Scahill, R., Harrison, G., Stevens, J., and Rudge, P. (2001). Long-term retrograde amnesia: The crucial role of the hippocampuus. *Neuropsychologia*, 39:151–172.

Coles, M. G. H. and Rugg, M. D. (1995). The ERP and cognitive psychology. In Rugg, M. D. and Coles, M. G. H., editors, *Electrophysiology of Mind: Event-related potentials and cognition.* New York: Oxford University Press.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33:497–505.

Corkin, S. (2002). What's new with the amnesic patient H.M.? *Nature Reviews Neuroscience*, 3:153–160.

Cowell, R. A., Bussey, T. J., and Saksida, L. M. (2010). Components of recognition memory: Dissociable cognitive processes or just differences in representational complexity? *Hippocampus*, 20:1245–1262.

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., and Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, 125:159–180.

Creelman, C. D. (1965). Discriminability and scaling of linear extent. *Journal of Experimental Psychology*, 70:192–200.

Criss, A. H. and Shiffrin, R. M. (2004). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, 32:1284–1297.

Curran, H. V., Gardiner, J. M., Java, R. I., and Allen, D. (1993). Effects of lorazepam upon recollective experience in recognition memory. *Psychopharmacology*, 110:374–378.

Curran, T. (1999). The electrophysiology of incidental and intentional retrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia*, 37:771–785.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory and Cognition*, 28:923–938.

Curran, T. (2004). Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia*, 42:1088–1106.

Curran, T. and Cleary, A. M. (2003). Using ERPs to dissociate recollection from familiarity in picture recognition. *Cognitive Brain Research*, 15:191–205.

Curran, T., Tanaka, J. W., and Weiskopf, D. M. (2002). An electrophysiological comparison of visual categorization and recognition memory. *Cognitive, Affective & Behavioural Neuroscience*, 2:1–18.

Curran, T., Tepe, K. L., and Piatt, C. (2006). Erp explorations of dual processes in recognition memory. In Zimmer, H. D., Mecklinger, A., and Lindenberger, U., editors, *Binding in Human Memory: A Neurocognitive Approach*, pages 467–492. Oxford: Oxford University Press.

Cycowicz, Y. M. and Friedman, D. (2003). Source memory for the color of pictures: Event-related brain potentials (ERPs) reveal sensory-specific retrieval-related activity. *Psychophysiology*, 40:455–464.

Cycowicz, Y. M. and Friedman, D. (2007). Visual novel stimuli in an ERP novelty oddball paradigm: Effects of familiarity on repetition and recognition memory. *Psychophysiology*, 44:11–29.

Cycowicz, Y. M., Friedman, D., and Snodgrass, J. G. (2001). Remembering the color of objects: An ERP investigation of source memory. *Cerebral Cortex*, 11:322–334.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion Neurobiology*, 16:693–700.

Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: Distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences*, 100:2157–2162.

Davies, G. M. and Thompson, D. M. (1988). *Memory in context: Context in memory.* Oxford: John Wiley & Sons.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, 109:710–721.

DeCarlo, L. T. (2003). An application of signal detection theory with finite mixture distributions to source discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29:767–778.

Diana, R. A., Yonelinas, A. P., and Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: A three-component model. *Trends in Cognitive Sciences*, 11:379–386.

Diana, R. A., Yonelinas, A. P., and Ranganath, C. (2008). The effects of unitization on familiarity-based source memory: Testing a behavioral prediction

derived from neuroimaging data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34:730–740.

Dodson, C. S. and Johnson, M. K. (1996). Some problems with the process-dissociation approach to memory. *Journal of Experimental Psychology: General*, 125:181–194.

Donaldson, D. I., Allan, K., and Wilding, E. L. (2002). Fractionating episodic memory retrieval using event-related potentials. In Parker, A., Wilding, E. L., and Bussey, T., editors, *The Cognitive Neuroscience of Memory: Encoding and Retrieval.*, pages 39–58. Hove: Psychology Press.

Donaldson, D. I. and Rugg, M. D. (1998). Recognition memory for new associations: Electrophysiological evidence for the role of recollection. *Neuropsychologia*, 36:377–395.

Donaldson, D. I., Wheeler, M. E., and Peterson, S. E. (2010). Remember the source: Dissociating frontal and parietal contributions to episodic memory. *Journal of Cognitive Neuroscience*, 22:377–391.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory and Cognition*, 24:523–533.

Donders, F. C. (1868). Over de snelheid van psychishe processen. *Acta Psychologica*, 30:412–431. Reprint: Translated by Koster, W. G. (1969).

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, 111:524–542.

Düzel, E., Vargha-Khadem, F., Heinze, H. J., and Mishkin, M. (2001). Brain activity evidence for recognition without recollection after early hippocampal damage. *Proceedings of the National Academy of Sciences of the United States of America*, 98:8101–8106.

Egan, J. P. (1958). Recognition memory and the operating characteristic. Technical Report AFCRC-TN-58-51, AD-152650, Indiana University, Hearing and Communication Laboratory.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* New York: Academic Press.

Eichenbaum, H., Fortin, N. J., Sauvage, M. M., Robitsek, R. J., and Farovik,

A. (2010). An animal model of amnesia that uses receiver operating characteristics (ROC) analysis to distinguish recollection from familiarity deficits in recognition memory. *Neuropsychologia*, 48:2281–2289.

Eichenbaum, H., Sauvage, M. M., Fortin, N. J., and Yonelinas, A. P. (2008). ROCs in rats? Response to wixted and squire. *Learning and Memory*, 15:691–693.

Eichenbaum, H., Yonelinas, A. P., and Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, 30:123–152.

Eldridge, L. L., Engel, S. A., Zeineh, M. M., Bookheimer, S. Y., and Knowlton, B. J. (2005). A dissociation of encoding and retrieval processes in the human hippocampus. *Journal of Neuroscience*, 25:3280–3286.

Elfman, K. W., Parks, C. M., and Yonelinas, A. P. (2008). Testing a neurocomputational model of recollection, familiarity, and source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34:752–768.

Fagot, J. and Cook, R. G. (2006). Evidence for large long-term memory capacities in baboons and pigeons and its implications for learning and the evolution of cognition. *Proceedings of the National Academy of Sciences USA*, 103:17564–17567.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874.

Ford, J. H., Verfaellie, M., and Giovanello, K. S. (2010). Neural correlates of familiarity-based associative retrieval. *Neuropsychologia*, 48:3019–3025.

Fortin, N. J., Wright, S. P., and Eichenbaum, H. (2004). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature*, 431:188–191.

Friedman, D., Cycowicz, Y. M., and Bersick, M. (2005). The late negative episodic memory effect: The effect of recapitulating study details at test. *Brain Research*, 23:185–198.

Friston, K. J., Price, C. J., Fletcher, P., Moore, C., Frackowiak, R. S. J., and Dolan, R. J. (1996). The trouble with cognitive subtraction. *NeuroImage*, 4:97–104.

Galli, G. and Otten, L. J. (2011). Material-specific neural correlates of recollec-

tion: Objects, words, and faces. *Journal of Cognitive Neuroscience*, 23:1405–1418.

Gardiner, J. M. and Gregg, V. H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review*, 4:474–479.

Gardiner, J. M., Java, R. I., and Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology*, 50:114–122.

Giovanello, K. S., Keane, M. M., and Verfaellie, M. (2006). The contribution of familiarity to associative memory in amnesia. *Neuropsychologia*, 44:1859–1865.

Glanzer, M. and Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16:5–16.

Glanzer, M., Hilford, A., and Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30:1176–1195.

Glanzer, M., Kim, K., Hilford, A., and Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:500–513.

Glaser, E. M. and Ruchkin, D. (1976). *Principles of neurobiological signal analysis.* New York: Academic Press.

Gobet, F. (1998). Expert memory: A comparison of four theories. *Cognition*, 66:115–152.

Godden, D. and Baddeley, A. (1980). When does context influence recognition memory? *British Journal of Psychology*, 71:99–104.

Goodenough, D. J., Rossman, K., and Lusted, L. B. (1972). Radiographic applications of signal detection theory. *Radiology*, 105:199–200.

Graf, P. and Schacter, D. I. (1989). Unitization and grouping mediate dissociations in memory for new associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15:930–940.

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Gregg, V. H. and Gardiner, J. M. (1994). Recognition memory and awareness: A large effect of study-test modalities on 'know' responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology*, 6:137–147.

Greve, A., Donaldson, D. I., and van Rossum, M. C. W. (2010). A single trace dual-process model of episodic memory: A novel computational account of familiarity and recollection. *Hippocampus*, 20:235–251.

Greve, A., van Rossum, M. C. W., and Donaldson, D. I. (2007). Investigating the functional interaction between semantic and episodic memory: Convergent behavioral and electrophysiological evidence for the role of familiarity. *NeuroImage*, 34:801–814.

Gronlund, S. D. and Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:1355–1369.

Hamann, S. B. and Squire, L. R. (1997). Intact priming for novel perceptual representations in amnesia. *Journal of Cognitive Neuroscience*, 9:699–713.

Handy, T. C. (2005). Basic principles of ERP quantification. In Handy, T. C., editor, *Event-Related Potentials: A Methods Handbook.*, pages 33–55. Cambridge, MA: MIT Press.

Harlow, I. M., MacKenzie, G., and Donaldson, D. I. (2010). Familiarity for associations? A test of the domain dichotomy theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36:1381–1388.

Haskins, A. L., Yonelinas, A. P., Quamme, J. R., and Ranganath, C. (2008). Perirhinal cortex supports encoding and familiarity-based recognition of novel associations. *Neuron*, 59:554–560.

Hasselmo, M. E. and Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, 89:1–34.

Hautus, M. J., Macmillan, N. A., and Rotello, C. M. (2008). Toward a com-

plete decision model of item and source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15:889–905.

Healy, M. R., Light, L. L., and Chung, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31:768–788.

Henson, R. N. and Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, 20:1315–1326.

Henson, R. N. A. (2005). A mini-review of fMRI studies of human medial temporal lobe activity associated with recognition memory. *Quarterly Journal of Experimental Psychology*, 58:340–360.

Herron, J. E. and Rugg, M. D. (2003). Strategic influences on recollection in the exclusion task: Electrophysiological evidence. *Psychonomic Bulletin & Review*, 10:703–710.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the theory of neural computations.* Redwood City, CA: Addison-Wesley.

Hicks, J. L., Marsh, R. L., and Ritschel, L. (2002). The role of recollection and partial information in source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28:503–508.

Hilford, A., Glanzer, M., Kim, K., and DeCarlo, L. T. (2002). Regularities of source recognition: ROC analysis. *Journal of Experimental Psychology: General*, 131:494–510.

Hintzman, D. I., Caulton, D. A., and T, C. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:275–289.

Hirshman, E., Fisher, J., Henthorn, T., Arndt, J., and Passanante, A. (2002). Midazolam amnesia and dual-process models of the word-frequency mirror effect. *Journal of Memory and Language*, 47:499–516.

Hirshman, E. and Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory & Cognition*, 25:345–351.

Hockley, W. E. and Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory and Cognition*, 27:657–664.

Holdstock, J. S., Gurnikov, S. A., Gaffan, D., and Mayes, A. R. (2000). Perceptual and mnemonic matching-to-sample in humans: Contributions of the hippocampus, perirhinal and other medial temporal lobe cortices. *Cortex*, 36:301–322.

Holdstock, J. S., Mayes, A. R., Gong, Q. Y., Roberts, N., and Kapur, N. (2005). Item recognition is less impaired than recall and associative recognition in a patient with selective hippocampal damage. *Hippocampus*, 15:203–215.

Holdstock, J. S., Parslow, D. M., Morris, R. G., Fleminger, S., Abrahams, S., Denby, C., Montaldi, D., and Mayes, A. R. (2002). Two case studies illustrating how relatively selective hippocampal lesions in humans can have quite different effects on memory. *Hippocampus*, 18:679–691.

Howard, M. W., Bessette-Symons, B., Zhang, Y., and Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: Evidence from modeling and receiver operating characteristic curves. *Psychology and Aging*, 21:96–106.

Jackson, III, O. and Schacter, D. L. (2004). Encoding activity in anterior medial temporal lobe supports subsequent associative recognition. *NeuroImage*, 21:456–462.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30:513–541.

Jacoby, L. L. (1998). Invariance in automatic influences of memory: Toward a user's guide for the process-dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24:3–26.

Jacoby, L. L. and Kelley, C. M. (1992). Unconscious influences of memory: Dissociations and automaticity. In Milner, A. D. and Rugg, M. D., editors, *The Neuropsychology of Consciousness. Foundations of Neuropsychology.*, pages 201–233. San Diego, CA: Academic Press.

Jennings, J. M. and Jacoby, L. L. (1997). Improving age-related deficits in recol-

lection: Application of an opposition procedure. *Brain and Cognition*, 35:403–406.

Jennings, J. R. and Wood, C. C. (1976). The epsilon-adjustment procedure for repeated-measures analyses of variance. *Psychophysiology*, 13:277–278.

Jäger, T., Mecklinger, A., and Kipp, K. H. (2006). Intra- and inter-item associations doubly dissociate the electrophysiological correlates of familiarity and recollection. *Neuron*, 52:535–545.

Jäger, T., Szabo, K., Griebe, M., Bäzner, H., J, M., and Hennerici, M. G. (2009). Selective disruption of hippocampus-mediated recognition memory processes after episodes of transient global amnesia. *Neuropsychologia*, 47:70–76.

Johansson, M. and Mecklinger, A. (2003). The late posterior negativity in ERP studies of episodic memory: Action monitoring and retrieval of attribute conjunctions. *Biological Psychology*, 64:91–117.

Johansson, M., Stenberg, G., Lindgren, M., and Rosén, I. (2002). Memory for perceived and imagined pictures: An event-related potential study. *Neuropsychologia*, 40:986–1002.

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114:3–28.

Joordens, S. and Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:1534–1555.

Joordens, S. and Merikle, P. M. (1993). Independence or redundancy? Two models of conscious and unconscious influences. *Journal of Experimental Psychology: General*, 122:462–467.

Kelley, R. and Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27:701–722.

Khoe, W., Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., and Knight, R. T. (2000). The contribution of recollection and familiarity to yes-no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia*, 38:1333–1341.

Kinoshita, S. (1997). Masked target priming effects on feeling-of-knowing and feeling-of-familiarity judgments. *Acta Psychologica*, 97:183–199.

Kirwan, C. B. and Stark, C. E. (2004). Medial temporal lobe activation during encoding and retrieval of novel face-name pairs. *Hippocampus*, 14:919–930.

Kirwan, C. B., Wixted, J. T., and Squire, L. R. (2010). A demonstration that the hippocampus supports both recollection and familiarity. *Proceedings of the National Academy of Sciences of the United States of America*, 107:344–348.

Knowlton, B. J. and Squire, L. R. (1995). Remembering and knowing: Two different expressions of declarative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21:699–710.

Koen, J. D. and Yonelinas, A. P. (2010). Memory variability is due to the contribution of recollection and familiarity, not to encoding variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36:1536–1542.

Kopelman, M. D. (1989). Remote and autobiographical memory, temporal context memory and frontal atrophy in Korsakoff and Alzheimer patients. *Neuropsychologia*, 27:437–460.

Kounios, J., Bachman, P., Casasanto, D., Grossman, M., Smith, R. W., and Yang, W. (2003). Novel concepts mediate word retrieval from human episodic associative memory: Evidence from event-related potentials. *Neuroscience Letters*, 345:157–160.

Kuo, T. and Van Petten, C. (2006). Prefrontal engagement during source memory retrieval depends on the prior encoding task. *Journal of Cognitive Neuroscience*, 18:1133–1146.

Kutas, M. and Dale, A. (1997). Electrical and magnetic readings of mental functions. In Rugg, M. D., editor, *Electrical and magnetic readings of mental functions.*, pages 197–241. Cambridge, MA: MIT Press.

Lecompte, D. C. (1995). Recollective experience in the revelation effect: Separating the contributions of recollection and familiarity. *Memory and Cognition*, 23:324–334.

Leynes, P. A. and Phillips, M. C. (2008). Event related potential (ERP) evidence

for varied recollection during source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34:741–751.

Li, J., Morcom, A. M., and Rugg, M. D. (2004). The effects of age on the neural correlates of successful episodic retrieval: An ERP study. *Cognitive, Affective & Behavioural Neuroscience*, 4:279–293.

Lins, O. G., Picton, T., Berg, P., and Scherg, M. (1993). Ocular artifacts in EEG and event-related potentials I: Scalp topography. *Brain Topography*, 6:51–63.

Lockhart, R. S. and Murdock, Jr, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, 74:100–109.

Lodish, H., Berk, A., Zipursky, L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). *Molecular Cell Biology.* New York: W. H. Freeman & Co Ltd.

Luck, S. J. (2005). *An introduction to the event-related potential technique.* Cambridge, MA: MIT Press.

Luck, S. J. and Zhang, W. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science*, 20:423–428.

MacAndrew, S. B. G., Jones, G. V., and Mayes, A. R. (1994). No selective deficit in recall in amnesia. *Memory*, 2:241–254.

Macho, S. (2002). Cognitive modeling with spreadsheets. *Behaviour Research Methods, Instruments & Computers*, 34:19–36.

MacKenzie, G. and Donaldson, D. I. (2007). Dissociating recollection from familiarity: Electrophysiological evidence that familiarity for faces is associated with a posterior old/new effect. *Neuroimage*, 36:454–463.

Macmillan, N. A. and Creelman, C. D. (2005). *Detection theory: A user's guide.* New York: Cambridge University Press, 2nd edition.

Macmillan, N. A., Rotello, C. M., and Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychonomics*, 66:406–421.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28:380–387.

Mandler, G. (1979). Organization and repetition: Organizational principles with

special reference to rote learning. In Nilsson, L. G., editor, *Perspectives on memory research.*, pages 293–327. Hillsdale, NJ: Erlbaum.

Mandler, G. (1980). Recognizing: The judgement of previous occurrence. *Psychological Review*, 87:252–271.

Mandler, G., Graf, P., and Kraft, D. (1986). Activation and elaboration effects in recognition and word priming. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 38:645–662.

Manns, J. R., Hopkins, R. O., Reed, J. M., Kitchener, E., and Squire, L. R. (2003). Recognition memory and the human hippocampus. *Neuron*, 37:171–180.

Manns, J. R. and Squire, L. R. (1999). Impaired recognition memory on the doors and people test after damage limited to the hippocampal region. *Hippocampus*, 9:495–499.

Marcum, J. I. (1947). A statistical theory of target detection by pulsed radar. Research Memorandum AD 101287, RAND Corporation.

Mayes, A. R., Holdstock, J. S., Isaac, C. L., Hunkin, N., and Roberts, N. (2002). Relative sparing of item recognition memory in a patient with adult-onset damage limited to the hippocampus. *Hippocampus*, 12:325–340.

Mayes, A. R., Holdstock, J. S., Isaac, C. L., Montaldi, D., Grigor, J., Gummer, A., Cariga, P., Downes, J. J., Tsivilis, D., Gaffan, D., Gong, Q., and Norman, K. A. (2004). Associative recognition in a patient with selective hippocampal lesions and relatively normal item recognition. *Hippocampus*, 14:763–784.

Mayes, A. R., Isaac, C. L., Holdstock, J. S., Hunkin, N. M., Montaldi, D., Downes, J. J., MacDonald, C., Cezayirli, E., and Roberts, J. N. (2001). Memory for single items, word pairs, and temporal order of different kinds in a patient with selective hippocampal lesions. *Cognitive Neuropsychology*, 18:97–123.

Mayes, A. R., Montaldi, D., and Migo, E. (2007). Associative memory and the medial temporal lobes. *Trends in Cognitive Sciences*, 11:126–135.

McCarthy, G. and Wood, C. C. (1985). Scalp distributions of event-related potentials: An ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology*, 62:203–208.

Mecklinger, A. (2000). Interfacing mind and brain: A neurocognitive model of recognition memory. *Psychophysiology*, 37:565–582.

Mickes, L., Johnson, E., and Wixted, J. T. (2010). Continuous recollection vs. unitized familiarity in associative recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36:843–863.

Mickes, L., Wais, P. E., and Wixted, J. T. (2009). Recollection is a continuous process: Implications for dual-process theories of recognition memory. *Psychological Science*, 20:509–515.

Milner, B., Corkin, S., and Teuber, H. L. (1968). Further analysis of the hippocampal amnesic syndrome: 14-year follow up study of H.M. *Neuropsychologia*, 6:215–234.

Montaldi, D. and Mayes, A. R. (2010). The role of recollection and familiarity in the functional differentiation of the medial temporal lobes. *Hippocampus*, 20:1291–1314.

Montaldi, D., Spencer, T. J., Roberts, N., and Mayes, A. R. (2006). The neural system that mediates familiarity memory. *Hippocampus*, 16:504–520.

Morris, C. D., Bransford, J. D., and Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behaviour*, 16:519–533.

Moscovitch, M., Vriezen, E., and Goshen-Gottstein, Y. (1993). Implicit tests of memory in patients with focal lesions or degenerative brain disorders. In Boller, F. and Spinnler, H., editors, *The Handbook of Neuropsychology*, volume 8. Amsterdam, the Netherlands: Elsevier.

Murdock, B. B. (1974). *Human memory: The theory and data.* Hillsdale, NJ: Erlbaum.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:1170–1187.

Nessler, D., Mecklinger, A., and Penney, T. B. (2001). Event related brain potentials and illusory memories: The effects of differential encoding. *Cognitive Brain Research*, 10:283–301.

Niedermeyer, E. and Lopes da Silva, F. H. (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.* Philadelphia: Lippincot Williams & Wilkins.

Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: Revisiting the complementary learning systems model. *Hippocampus*, 20:1217–1227.

Norman, K. A. and O'Reilly, R. A. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning-systems approach. *Psychological Review*, 110:611–646.

Olichney, J., Van Petten, C., Paller, K. A., Salmon, D., Iragui, V., and Kutas, M. (2000). Word repetition in amnesia: Electrophysiological evidence of spared and impaired memory. *Brain*, 123:1948–1963.

Onyper, S. V., Zhang, Y. X., and Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 139:341–364.

Opitz, B. and Cornell, S. (2006). Contribution of familiarity and recollection to associative recognition memory: Insights from event-related potentials. *Journal of Cognitive Neuroscience*, 18:1595–1605.

Paller, K. A. and Kutas, M. (1992). Brain potentials during memory retrieval provide neurophysiological support for the distinction between conscious recollection and priming. *Journal of Cognitive Neuroscience*, 4:375–391.

Parks, C. M. and Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on wixted. *Psychological Review*, 114:188–202.

Parks, T. E. (1966). Signal detectability theory of recognition memory performance. *Psychological Review*, 73:44–58.

Pastore, R. E., Crawley, E. J., Berens, M. S., and Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, 10:556–569.

Perfect, T. J., Mayes, A. R., Downes, J. J., and Eijk, R. V. (1996). Does context discriminate recollection from familiarity in recognition memory? *Quarterly Journal of Experimental Psychology*, 49:797–813.

Perry, N. W. (1966). Signal versus noise in evoked potential. *Science*, 153:1022.

Peters, J., Thoma, P., Koch, B., Schwarz, M., and Daum, I. (2009). Impairment of verbal recollection following ischemic damage to the right anterior hippocampus. *Cortex*, 45:592–601.

Ploran, E. P., Nelson, S. M., Velanova, K., Donaldson, D. I., Peterson, S. E., and Wheeler, M. E. (2007). Evidence accumulation and the moment of recognition: Dissociating perceptual recognition processes using fMRI. *Journal of Neuroscience*, 27:11912–11924.

Postma, A. (1999). The influence of decision criteria upon remembering and knowing in recognition memory. *Acta Psychologica*, 103:65–76.

Quamme, J. R., Yonelinas, A. P., and Norman, K. A. (2007). Effect of unititization on associative recognition in amnesia. *Hippocampus*, 17:192–200.

Quamme, J. R., Yonelinas, A. P., Widaman, K. F., Kroll, N. E. A., and Sauve, M. J. (2004). Recall and recognition in mild hypoxia: Using covariance structural modeling to test competing theories of explicit memory. *Neuropsychologia*, 42:672–691.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory and Cognition*, 21:89–102.

Rajaram, S. (1996). Perceptual effects on remembering: Recollective processes in picture recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:365–377.

Ranganath, C. (2010). A unified framework for the functional organization of the medial temporal lobes and the phenomenology of episodic memory. *Hippocampus*, 20:1263–1290.

Ranganath, C. and Paller, K. A. (2000). Neural correlates of memory retrieval and evaluation. *Cognitive Brain Research*, 9:209–222.

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., and D'Esposito, M. (2003). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia*, 42:2–13.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85:59–108.

Ratcliff, R., McKoon, G., and Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:763–785.

Ratcliff, R., Sheu, C., and Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, 99:518–535.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M., Angstadt, P., and Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:294–320.

Reed, J. M. and Squire, L. R. (1997). Impaired recognition memory in patients with lesions limited to the hippocampal formation. *Behavioural Neuroscience*, 111:667–675.

Rempel-Clower, N. L., Zola, S. M., Squire, L. R., and Amaral, D. G. (1996). Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. *Journal of Neuroscience*, 16:5233–5255.

Rhodes, S. M. and Donaldson, D. I. (2007). Electrophysiological evidence for the influence of unitization on the processes engaged during episodic retrieval: Enhancing familiarity based remembering. *Neuropsychologia*, 45:412–424.

Rhodes, S. M. and Donaldson, D. I. (2008). Electrophysiological evidence for the effect of interactive imagery on episodic memory: Encouraging familiarity for non-unitized stimuli during associative recognition. *NeuroImage*, 39:873–884.

Roediger, H. L. and Bergman, E. T. (1998). The controvery over recovered memories. *Psychology, Public Policy & Law*, 4:1091–1109.

Roediger, H. L. and McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21:803–814.

Rotello, C. M. and Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, 28:907–922.

Rotello, C. M., Macmillan, N. A., and Reeder, J. A. (2004). Sum-difference the-

ory of remembering and knowing: A two-dimensional signal detection model. *Psychological Review*, 111:588–616.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., and Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, 12:865–873.

Rouder, J. N., Pratte, M. S., and Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin & Review*, 17:427–435.

Rugg, M. D. and Allan, K. (2000). Event-related potential studies of memory. In Tulving, E. and Craik, F. I. M., editors, *Oxford Handbook of Memory.*, pages 521–537. London: Oxford University Press.

Rugg, M. D. and Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, 11:251–257.

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., and Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature*, 392:595–598.

Rugg, M. D., Otten, L. J., and Henson, R. N. A. (2002). The neural basis of episodic memory: Evidence from functional neuroimaging. *The Philosophical Transactions of the Royal Society, Series B*, 357:1097–1110.

Rugg, M. D. and Yonelinas, A. P. (2003). Human recognition memory: A cognitive neuroscience perspective. *Trends in Cognitive Sciences*, 7:313–319.

Sauvage, M. M., Beer, Z., and Eichenbaum, H. (2010). Recognition memory: Adding a response deadline eliminates recollection but spares familiarity. *Learning and Memory*, 17:104–108.

Sauvage, M. M., Fortin, N. J., Owens, C. B., Yonelinas, A. P., and Eichenbaum, H. (2008). Recognition memory: Opposite effects of hippocampal damage on recollection and familiarity. *Nature Neuroscience*, 11:16–18.

Schacter, D. L., Savage, C. R., Alpert, N. M., Rauch, S. L., and Albert, M. S. (1996). The role of hippocampus and frontal cortex in age-related memory changes: A PET study. *NeuroReport*, 7:1165–1169.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood

estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610.

Senkfor, A. J., Van Petten, C., and Kutas, M. (2002). Episodic action memory for real objects: An ERP investigation with perform, watch, and imagine action encoding tasks versus a non-action encoding task. *Journal of Cognitive Neuroscience*, 14:402–419.

Sherman, S. J., Arri, A., Hasselmo, M. E., Stern, C. E., and Howard, M. W. (2003). Scopalomine impairs human recognition memory: Data and modeling. *Behavioural Neuroscience*, 114:526–539.

Shimamura, A. P. (2010). Hierarchical relational binding in the medial temporal lobe: The strong get stronger. *Hippocampus*, 20:1206–1216.

Shimamura, A. P. and Wickens, T. D. (2009). Superadditivity memory strength for item and source recognition: The role of hierarchical relational binding in the medial temporal lobe. *Psychological Review*, 116:1–19.

Slotnick, S. D. (2010). Remember source memory ROCs indicate recollection is a continuous process. *Memory*, 18:27–39.

Slotnick, S. D. and Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, 33:151–170.

Smith, M. E. (1993). Neurophysiological manifestations of recollective experience during recognition memory judgements. *Journal of Cognitive Neuroscience*, 5:1–13.

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry: The principles and practice of statistics in biological research.*, volume 3. New York: W. H. Freeman and Co.

Speer, N. K. and Curran, T. (2007). ERP correlates of familiarity and recollection processes in visual associative recognition. *Brain Research*, 1174:97–109.

Squire, L. R. and Knowlton, B. J. (1995). Memory, hippocampus, and brain systems. In Gazzaniga, M. S., editor, *The Cognitive Neurosciences.*, pages 825–837. Cambridge, MA: MIT Press.

Squire, L. R., Wixted, J. T., and Clark, R. E. (2007). Recognition memory and

the medial temporal lobe: A new perspective. *Nature Reviews Neuroscience*, 8:872–883.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, 25:207–222.

Stark, C. E. and Squire, L. R. (2000). Recognition memory and familiarity judgments in severe amnesia: no evidence for a contribution of repetition priming. *Behavioural Neuroscience*, 114:459–467.

Stark, C. E. L., Bayley, P. J., and Squire, L. R. (2002). Recognition memory for single items and for associations is similarly impaired following damage to the hippocampal region. *Learning and Memory*, 9:238–242.

Stark, C. E. L. and Squire, L. R. (2003). Hippocampal damage equally impairs memory for single items and memory for conjunctions. *Hippocampus*, 13:239–250.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, 30:276–315.

Strack, F. and Foerster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science*, 6:352–358.

Swets, J. A., Tanner, W. P., and Birdsall, Jr, T. G. (1961). Decision processes in perception. *Psychological Review*, 68:301–340.

Temple, C. M. and Richardson, P. (2004). Developmental amnesia: a new pattern of dissociation with intact episodic memory. *Neuropsychologia*, 42:764–781.

Toth, J. P. (1996). Conceptual automaticity in recognition memory: Levels-of-processing effects on familiarity. *Canadian Journal of Experimental Psychology*, 50:123–138.

Treves, A. and Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4:374–391.

Trott, C. T., Friedman, D., Ritter, W., Fabiani, M., and Snodgrass, J. G. (1999). Episodic priming and memory for temporal source: Event-related potentials reveal age-related differences in prefrontal functioning. *Psychology and Aging*, 14:390–413.

Troyer, A. K., Winocur, G., Craik, F. I. M., and Moscovitch, M. (1999). Source memory and divided attention: Reciprocal costs to primary and secondary tasks. *Neuropsychology*, 13:467–474.

Tsivilis, D., Vann, S. D., Denby, C., Roberts, J. N., Mayes, A. R., Montaldi, D., and Aggleton, J. P. (2008). A disproportionate role for the fornix and mammillary bodies in recall versus recognition memory. *Nature Neuroscience*, 11:834–842.

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E. and Donaldson, W., editors, *Organization of Memory.*, pages 381–403. New York: Academic Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26:1–12.

Turriziani, P., Fadda, L., Clatagirone, C., and Carlesimo, G. A. (2004). Recognition memory for single items and for associations in amnesic patients. *Neuropsychologia*, 42:426–433.

Urbach, T. P. and Kutas, M. (2002). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology*, 39:791–808.

Vann, S. D., Tsivilis, D., Denby, C. E., Quamme, J. R., Yonelinas, A. P., and Aggleton, J. P. (2009). Impaired recollection but spared familiarity in patients with extended hippocampal system damage revealed by 3 convergent methods. *Proceedings of the National Academy of Sciences of the United States of America*, 106:5442–5447.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., and Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, 277:376–380.

Verfaellie, M., Koseff, P., and Alexander, M. P. (2000). Acquisition of novel semantic information in amnesia: effects of lesion location. *Neuropsychologia*, 38:484–492.

Verfaellie, M. and Treadwell, J. R. (1993). Status of recognition memory in amnesia. *Neuropsychology*, 7:5–13.

Vilberg, K. L., Moosavi, R. F., and Rugg, M. D. (2006). The relationship be-

tween electrophysiological correlates of recollection and amount of information retrieved. *Brain Research*, 1122:161–170.

Voss, J. L., Hauner, K. K., and Paller, K. A. (2009). Establishing a relationship between activity reduction in human perirhinal cortex and priming. *Hippocampus*, 19:773–778.

Voss, J. L. and Paller, K. A. (2006). Fluent conceptual priming and explicit memory for faces are electrophysiologically distinct. *Journal of Neuroscience*, 26:926–933.

Wagner, A. D., Gabrieli, J. D. E., and Verfaellie, M. (1997). Dissociations between familiarity processes in explicit recognition and implicit perceptual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:305–323.

Wais, P. E., Mickes, L., and Wixted, J. T. (2008). Remember/know judgments probe degrees of recollection. *Journal of Cognitive Neuroscience*, 20:400–405.

Wais, P. E., Wixted, J. T., Hopkins, R. O., and Squire, L. R. (2006). The hippocampus supports both the recollection and the familiarity components of recognition memory. *Neuron*, 49:459–466.

Westerberg, C. E., Paller, K. A., Weintraub, S., Mesulam, M. M., Holdstock, J. S., Mayes, A. R., and Reber, P. J. (2006). When memory does not fail: Familiarity-based recognition in mild cognitive impairment and Alzheimer's disease. *Neuropsychology*, 20:193–205.

Westerman, D. L. (2001). The role of familiarity in item recognition, associative recognition, and plurality recognition on self-paced and speeded tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27:723–732.

Wickelgren, W. A. and Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology*, 2:316–347.

Wickens, T. D. (2002). *Elementary signal detection theory.* New York: Oxford University Press.

Wilding, E. L. (1999). Separating retrieval strategies from retrieval success: An event-related potential study of source memory. *Neuropsychologia*, 37:441–454.

Wilding, E. L. (2000). In what way does the parietal ERP old/new effect index recollection? *International Journal of Psychophysiology*, 35:81–87.

Wilding, E. L. (2006). On the practice of rescaling scalp-recorded electrophysiological data. *Biological Psychology*, 72:325–332.

Wilding, E. L. and Rugg, M. D. (1997). An event-related potential study of memory for words spoken aloud or heard. *Neuropsychologia*, 9:1185–1195.

Wixted, J. T. (2007a). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, 114:152–176.

Wixted, J. T. (2007b). Spotlighting the probative findings: Reply to Parks and Yonelinas (2007). *Psychological Review*, 114:203–209.

Wixted, J. T., Mickes, L., and Squire, L. R. (2010). Measuring recollection and familiarity in the medial temporal lobe. *Hippocampus*, 20:1195–1205.

Wixted, J. T. and Squire, L. R. (2008). Constructing receiver operating characteristics (ROCs) with experimental animals: Cautionary notes. *Learning and Memory*, 15:687–690.

Wixted, J. T. and Stretch, V. (2004). In defense of the signal detection interpretation of Remember/Know judgments. *Psychonomic Bulletin & Review*, 11:616–641.

Wolk, D. A., Schacter, D. L., Lygizos, M., Mandu Sen, N., Chong, H., Holcolm, P. J., Daffner, K. R., and Budson, A. E. (2007). ERP correlates of remember/know decisions: Association with the late posterior negativity. *Biological Psychology*, 75:131–135.

Wood, C. C. (1987). Generators of event-related potentials. In Halliday, A. M., Bulter, S. R., and Paul, R., editors, *A Textbook of Clinical Neurophysiology.*, pages 535–567. John Wiley & Sons Ltd.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92:937–950.

Yick, Y. Y. and Wilding, E. L. (2008). Material- specific neural correlates of memory retrieval. *Neuro Report*, 19:1463–1467.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition mem-

ory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:1341–1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, 25:747–763.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:1415–1434.

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 cs of recognition memory. *Journal of Experimental Psychology: General*, 130:361–379.

Yonelinas, A. P. (2002a). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46:441–517.

Yonelinas, A. P. (2002b). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46:441–517.

Yonelinas, A. P., Aly, M., Wang, W.-C., and Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, 20:1178–1194.

Yonelinas, A. P. and Jacoby, L. L. (1995). The relation between remembering and knowing as a basis for recognition: Effects of size congruency. *Journal of Memory and Language*, 34:622–643.

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I., Lazzara, M., and Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember-know, process dissociation, and receiver operating characteristic data. *Neuropsychology*, 12:323–339.

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I. G., and Soltani, M. (1999). Recognition memory for faces: When familiarity supports associative recognition judgments. *Psychonomic Bulletin & Review*, 6:654–661.

Yonelinas, A. P. and Parks, C. M. (2007). Receiver operating characteristics

(ROCs) in recognition memory: A review. *Psychological Bulletin*, 133:800–832.

Yovel, G. and Paller, K. A. (2004). The neural basis of the butcher-on-the-bus phenomenon: When a face seems familiar but is not remembered. *Neuroimage*, 21:789–800.

Zola-Morgan, S. and Squire, L. R. (1986). Memory impairment in monkeys following lesions limited to the hippocampus. *Behavioural Neuroscience*, 100:155–160.